DOCUMENT RESUME

ED 251 512 TM 850 016

AUTHOR	Gialluca, Kathleen A.; And Others
TITLE	Methods for Equating Mental Tests. Interim Report for Period March 1982-October 1984.
INSTITUTION	Assessment Systems Corp., St. Paul, Minn.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Tex. Manpower and Personnel Div.
REPORT NO	AFHRL-TR-84-35
PUB DATE	Nov 84
CONTRACT	F41689-82-C-0023
NOTE	163p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Adults; Cognitive Ability; *Cognitive Tests; Comparative Analysis; Difficulty Level; *Equated Scores; *Latent Trait Theory; Research Design; Sample Size; Test Items; Test Length; *Test Theory; *True Scores
IDENTIFIERS	*Equipercentile Equating; *Linear Equating Method; Vertical Equating

ABSTRACT

In this study, simulated and actual Air Force test data were used to compare the different procedures for equating mental tests: conventional (equipercentile and linear), Item Response Theory (IRT), and strong true-score theory (STST); data collection designs used were single-group, equivalent-groups, and anchor test. Equating transformations were evaluated by comparing equated scores with true/observed scores along with bias and root-mean-squared-error indices. Specific testing-condition manipulations included variations in test length, item difficulty, sample size, and examinee ability distributions. The study found that parallel subtests were best equated using the simple conventional methods; nonparallel subtests, on the other hand, were best equated with the more complex IRT and STST methods. There were few differences among the data collection designs when they were applied to samples of equivalent ability levels; the anchor-test design was essential for equating subtests using nonequivalent examinee groups. There was little advantage to be gained by increasing the sample size from 1,000 to 2,400 examinees. Equating accuracy was not markedly affected when subtest length was doubled, nor did it matter whether easy or difficult subtests were equated. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AFHRL-TR-84-35

ED251512

AIR FORCE

HUMAN RESOURCES

METHODS FOR EQUATING MENTAL TESTS

By

Kathleen A. Gialluca
Leslie I. Crichton
C. David Vale

Assessment Systems Corporation
2233 University Avenue, Suite 310
St. Paul, Minnesota 55114

Malcolm James Ree

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5000

November 1984

Interim Report for Period March 1982 - October 1984

Approved for public release; distribution unlimited.

LABORATORY

TM 850 012

AIR FORCE SYSTEMS COMMAND

BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Manpower and Personnel Division

ANTHONY F. BRONZO, JR., Colonel, USAF
Commander

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFHRL-TR-84-35

6a. NAME OF PERFORMING ORGANIZATION
Assessment Systems Corp.

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Manpower and Personnel Division

6c. ADDRESS *(City, State, and ZIP Code)*
2233 University Avenue, Suite 310
St Paul, Minnesota 55414

7b. ADDRESS *(City, State, and ZIP Code)*
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5000

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Air Force Human Resources Laboratory

8b. OFFICE SYMBOL *(If applicable)*
HQ AFHRL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F41689-82-C-0023

8c. ADDRESS *(City, State, and ZIP Code)*
Brooks Air Force Base, Texas 78235-5000

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
62703F	7719	18	28

11. TITLE *(Include Security Classification)*
Methods for Equating Mental Tests

12. PERSONAL AUTHOR(S)
Gialluca, Kathleen A.; Crichton, Leslie I.; Vale. C. David; Ree, Malcolm J.

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM Mar 82 TO Oct 84

14. DATE OF REPORT *(Year, Month, Day)*
November 1984

15. PAGE COUNT
164

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
ASVAB
equating
item response theory
strong true score theory
tests and measurement

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The technology for test equating has arisen from the need to make new tests comparable to old ones. The equating of military tests has two objectives: (a) to make scores on different tests forms and on different composites of test forms comparable, and at the same time (b) to solve the norming problem by relating all scores on new tests and composites back to a large sample of talent indicative of an anticipated population of military enlistees. In this study, simulated and actual Air Force test data were used to compare the different procedures for equating mental tests and to delineate those conditions under which each equating procedure performed best. Specific testing-condition manipulations included variations in test length, item difficulty, sample size, and examinee ability distributions. Coventional (equipercentile and linear), Item Response Theory (IRT), and Strong True Score Theory (STST) methods comprised the equating procedures that were studied. The data collection designs that were used included the single-group, equivalent-groups, and anchor-test designs. Equating transformations were evaluated by comparing equated scores with true scores using bias and root-mean-squared-error indices.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Nancy A. Perrigo
Chief, STINFO Office

22b. TELEPHONE *(Include Area Code)*
(512) 536-3877

22c. OFFICE SYMBOL
AFHRL/TSR

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

Item 19. (Continued)

The present study found that parallel subtests were best equated using the simple conventional equating methods; nonparallel subtests, on the other hand, were best equated with the more complex IRT and STST methods. There were few differences among the data collection designs when they were applied to examinee examples of equivalent ability levels; the anchor-test design were essential for equating subtests using nonequivalent examinee groups. There was little advantage to be gained by increasing the sample size from 1,000 to 1,400 examinees. Equating accuracy was not markedly affected when subtest length was doubled, nor did it matter whether easy or difficult subtests were equated. Only IRT and STST methods were appropriate for equating subtests vertically across different levels.

Parallel composites were best equated by forming composites from subtests that had first been individually equated using IRT or STST methods. Nonparallel composites, on the other hand, were equated well only when STST procedures were applied to the composite score itself. No clear recommendations can be made regarding the choice of a data collection design of sample size for equating test composites, since no consistent differences among designs and sizes were noted. Vertical equating of composites is not recommended.

SUMMARY

The technology for test equating has arisen from the need to make new tests comparable to old ones. Equating of military tests has two objectives (1) to make scores on different test forms and composites of test forms comparable, and at the same time (2) to solve the norming problem by relating all scores on new tests and composites back to a large sample of talent indicative of an anticipated mobilization population. In this study, simulated and actual Air Force test data were used to compare the different procedures for equating mental tests and delineate those testing conditions under which each equating procedure performed best. Specific testing-condition manipulations included variations in test length, item difficulty, sample size, and examinee ability distributions. Equating procedures studied included conventional (equipercentile and linear), Item Response Theory (IRT), and strong true-score theory (STST); data collection designs used were single-group, equivalent-groups, and anchor-test. Equating transformations were evaluated by comparing equated scores with true/observed scores along with bias and root-mean-squared-error indices.

The study found that parallel subtests were best equated using the simple conventional methods; nonparallel subtests, on the other hand, were best equated with the more complex IRT and STST methods. There were few differences among the data collection designs when they were applied to samples of equivalent ability levels; the anchor-test design was essential for equating subtests using nonequivalent examinee groups. There was little advantage to be gained by increasing the sample size from 1,000 to 2,400 examinees. Equating accuracy was not markedly affected when subtest length was doubled, nor did it matter whether easy or difficult subtests were equated.

PREFACE

The studies presented in this report were accomplished as part of Project 7719, Force Acquisition and Distribution Systems. It is one in a series on the equating of tests and test items. The effort represents the concern of this Laboratory for maintaining and advancing the state-of-the-art in the applications of test and measurement theory.

ACKNOWLEDGEMENTS

The authors are indebted to R. S. Massar of U. S. MEPCOM, J. B. Sympson of NPRDC, and D. J. Weiss and D. S. Suhadolnik of the University of Minnesota for providing some of the data upon which these computer simulations were modeled.

TABLE OF CONTENTS

INTRODUCTION 13

Equating Needs in Educational Testing. 14

Educational Decisions. 14
Scale Meaningfulness 14

Equating Needs in Military Testing 15

History of Air Force Testing 15
Airman Classification Battery. 15
AC-1A. 15
AC-1B. 16
AC-2A. 16
Airman Qualifying Examinations 17
AQE-D. 17
AQE-F. 17
AQE-62 17
AQE-64 17
AQE-66 18
AQE-J. 18
Armed Services Vocational Aptitude Battery 18

Military vs. Educational Testing 19

OVERVIEW OF EQUATING MODELS. 20

Calibration vs. Equating: A Clarification of Terminology. . . . 20

Calibration. 20
Equating 20
Operational Definition of Equating 21

Data Collection Designs. 22

Equating Transformations 23

Conventional Equating. 23
Equipercentile Equating. 24
Raw transformation 24
Smoothing. 24
Linear Equating. 26
Problems with Conventional Equating Methods. 26

Item Response Theory . . . 27
True-Score Equating. . . . 27
Observed-Score Equating. . . . 28
Item Parameter Estimates . . . 28
Strong True-Score Theory . . . 29

EVALUATION OF EQUATING METHODS: A REVIEW. . . . 31

Previous Research. . . . 31

Conventional Equating Methods. . . . 31
Regression . . . 31
Linear . . . 33
Equipercentile . . . 33
Yen. . . . 33
Slinde-Linn. . . . 34
Comparisons Among Conventional Methods . . . 35
Bianchini-Loret. . . . 35
Stock-Kagan-Van Wagenen. . . . 35
Lord . . . 36
Summary. . . . 37
IRT Equating Methods . . . 37
One-Parameter Model. . . . 37
Slinde-Linn-Gustafsson . . . 37
Divgi. . . . 39
Loyd-Hoover. . . . 41
Guskey . . . 42
Holmes . . . 43
Summary. . . . 44
Three-Parameter Model. . . . 44
Cook-Eignor-Petersen . . . 44
Holmes . . . 45
Lord-Wingersky . . . 46
Summary. . . . 46
Comparisons Among IRT Methods. . . . 47
One- vs. three-parameter models . . . 47
One- vs. two- vs. three-parameter models . . . 47
Summary. . . . 48
Comparisons Between Conventional and IRT Methods . . . 48
Conventional vs. One-Parameter IRT. . . . 48
Rentz-Bashaw . . . 48
Beard-Pettie . . . 49
Golub-Smith. . . . 49
Conventional vs. Three-Parameter IRT . . . 50
Lord . . . 50
Marco. . . . 51
Bejar-Wingersky. . . . 51

- Modu 52
- Petersen-Cook-Stocking 52
- Hicks. 54
- Conventional vs. One- vs. Three-Parameter IRT. 55
  - Marco-Petersen-Stewart 55
  - Kolen-Whitney. 56
- Conventional vs. One- vs. Two- vs. Three-Parameter IRT . 57
  - Kolen. 57
  - Phillips 58
- Summary. 58

Relevance of Previous Research to Practical Equating Situations. 59

- Equating Parallel Tests. 59
  - Theoretically Appropriate Methods. 59
  - Previous Research. 59
  - Conclusions. 60
- Equating Nonparallel Tests of Equal Difficulty 60
  - Theoretically Appropriate Methods. 60
  - Previous Research. 61
  - Conclusions. 61
- Equating Tests of Different Difficulty 62
  - Theoretically Appropriate Methods. 62
  - Previous Research. 62
  - Conclusions. 62

The Criterion Problem. 63

- Previous Approaches. 63
  - Discrepancies Across Methods 63
  - Consistency and Stability Indices. 64
  - Equating a Test to Itself. 64
  - Discrepancies Between Observed Scores and Equated Scores 65
- Observed-Score vs. True-Score Equating 65
- A More Satisfactory Approach 66

Design Issues for a Study of Equating. 67

- Examinee Ability Distributions 67
- Test Structures. 67
- Sample Sizes 68
- Composites 68

METHOD 69

Project Overview 69

  - Sample Characteristics 70
  - Test Characteristics 71

Data Collection Designs. 72

Data-Generation Procedures 74
  - Examinee Characteristics 74
    - Specification of the Moments of the True-Ability Distributions. 74
      - Power-test abilities 74
      - Speeded-test abilities 75
      - Varying ability. 76
    - Specification of the Correlations Among True Abilities . 77
    - Sample Sizes and Combinations. 77
    - Generation of the True-Ability Distributions 77
  - Test Characteristics 79
    - Power Subtests 79
      - Test lengths 79
      - Specification of the true-item-parameter distributions. 79
      - Generation of the true-item-parameter distributions. 79
      - Assignment of items to individual test forms 82
    - Speeded Subtests 85
    - Composites 85
    - Selection Composite. 86
    - Anchor Tests 87
  - Generation of Item Responses 88
    - Power Subtests 88
    - Speeded Subtests 88
    - Adequacy of the Simulation Procedures. 90

Applications of Equating Transformations. 91

  - Linear Equating. 91
  - Equipercentile Equating. 91
    - Components of the Equating Procedure 92
      - Percentile tables. 92
      - Regression smoothing of percentile tables. 92
      - Spline smoothing of percentile tables. 92
      - Equating procedure 93
      - Regression smoothing of equating tables. 93
      - Spline smoothing of equating tables. 93
    - Comparison of the Smoothing Procedures 94
  - Item Response Theory Equating. 94
    - Item Calibration Program 95
    - Equating Procedure 96
  - Strong True-Score Theory Equating. 96
    - Estimating a Test's True-Score Distribution. 97

        Initial λ estimates. 98
        Refining the λ estimates 98
    Equating the Tests 98

Procedures for Equating Test Composites. 99

    Equating Composite Scores Directly 100
    Forming Composites of Equated Subtests 100
    Equating Composite Scores Indirectly Through the Subtests. .101

Evaluative Criteria. 103

Real-Data Application. 104

    Raw Data 104
    Data Collection Designs. 105
    Equating Transformations 105
    Evaluative Criteria. .105

RESULTS AND DISCUSSION 107

Choosing an Equipercentile Smoothing Method. 107

    Results. 107
    Discussion 108

Equating Individual Subtests 108

    Equating Methods 108
        Results. 108
        Discussion 109
    Data Collection Designs. 110
        Results. 110
        Discussion 111
    Sample Sizes 113
        Results. 113
        Discussion 114
    Test Lengths and Difficulties. 115
        Results. 115
        Discussion 118
    Ability Levels 118
        Results. 118
        Discussion 120

Equating Test Composites .120

    Equating Methods 120

Results. 120
Strong true-score theory 120
Equating composite scores directly 121
Forming composites of equated subtests 122
Equating composite scores indirectly through the subtests. 123
Discussion 124
Data Collection Designs. 124
Results. 124
Discussion 127
Sample Sizes 127
Results. 127
Discussion 130
Test Lengths and Difficulties. 130
Results. 130
Discussion 133

Real-Data Application. .133

Results. 133
Discussion 135

CONCLUSIONS AND RECOMMENDATIONS. 136

Individual Subtests. 136

Smoothing Methods. 136
Equating Methods 136
Data Collection Designs. 136
Sample Sizes 137
Test Lengths and Difficulties. 137

Composites 138

REFERENCES 140

APPENDIX A 149

LIST OF TABLES

1 Application of Equating Methods and Data Collection Designs to Subtests and Composites 70

2 Test-Form Characteristics 71

3 Test-Form Pairings 72

4 Combinations of Data Collection Designs and Sample Ability Distributions 73

5 Summary Statistics Used to Specify Multivariate Distribution of True Abilities 74

6 Distribution of Applicants Across AFQT Categories 76

7 Summary Statistics of Multivariate Distributions of True Abilities: Samples X, Y, Z, and W 78

8 Summary Statistics Used to Specify Multivariate Distributions of True Item Parameters: Subtests PC, AR, and WK 80

9 Summary Statistics of Multivariate Distributions of True Item Parameters: Subtest PC 81

10 Summary Statistics of Multivariate Distributions of True Item Parameters: Subtest AR 82

11 Summary Statistics of Multivariate Distributions of True Item Parameters: Subtest WK 83

12 Strategy for Assigning Items to Individual Subtest Forms 84

13 True Item Parameter Means for Each Test Form: Subtest PC . . . 85

14 True Item Parameter Means for Each Test Form: Subtest AR . . . 86

15 True Item Parameter Means for Each Test Form: Subtest WK . . . 87

16 Summary Statistics for Real and Simulated ASVAB Subtest Scores 90

17 Equipercentile Equating Smoothing Methods: True-Score Error Indices and Tally of "Best" Method for Subtest AR 95

18 Administration of Subtests to Examinee Subgroups: Creating Partial Data Sets 103

19 True-Score Error Indices for Equipercentile Smoothing Methods .107

20 True-Score Error Indices for Equating Subtests. 109

21 True-Score Error Indices for Equating Parallel Subtests Using Different Data Collection Designs 110

22 True-Score Error Indices for Equating Nonparallel Subtests Using DIfferent Data Collection Designs 111

23 True-Score Error Indices for Equating Parallel Subtests Using Various Sample Sizes. 113

24 True-Score Error Indices for Equating Nonparallel Subtests Using Various Sample Sizes. 114

25 True-Score Error Indices for Equating Parallel Subtests Using Various Levels of Test Difficulty and Length. 115

26 True-Score Error Indices for Equating Parallel Subtests Using Different Equating Methods and Various Levels of Test Length and Difficulty. 116

27 True-Score Error Indices for Equating Nonparallel Subtests Using Different Equating Methods and Various Levels of Test Length and Difficulty 117

28 True-Score Error Indices for Equating Across Ability Levels on Subtest PC 119

29 True-Score Error Indices for Equating Composite Scores Directly. 121

30 True-Score Error Indices for Forming Composites of Equated Subtests. 123

31 True-Score Error Indices for Equating Composites Indirectly Through the Subtests. 124

32 True-Score Error Indices for Equating Parallel Composites Using Different Data Collection Designs 125

33 True-Score Error Indices for Equating Nonparallel Composites Using Different Data Collection Designs 126

34 True-Score Error Indices for Equating Parallel Composites Using Various Sample Sizes. 128

35 True-Score Error Indices for Equating Nonparallel Composites Using Various Sample Sizes. .129

36 True-Score Error Indices for Equating Parallel Composites Using Various Levels of Composite Length and Difficulty 131

37 True-Score Error Indices for Equating Nonparallel Composites Using Various Levels of Composite Length and DIfficulty 132

38 Observed-Score Error Indices for Equating Methods and Data Collection Designs: Real-Data Verification 134

METHODS FOR EQUATING MENTAL TESTS

The concept of a standard is basic to all forms of measurement. Precise standards for many physical quantities such as the meter, for example, have been developed and universally accepted. Psychological measurement is somewhat less advanced. Although several psychological variables (e.g., IQ) have been quantified, these psychological characteristics are usually indexed by scores on a particular test rather than by a universally accepted standard.

One essential characteristic of a standard of measurement is its invariance. A standard that deteriorates and changes through use or storage is not a satisfactory standard. In the physical sciences, for example, a meter was originally defined as the length of a metal bar stored under ideal conditions at the International Bureau of Weights and Measures in France. Because this ultimately proved to be an unacceptable standard (both because of inaccuracy and impermanence), it was replaced by a specific number of wavelengths of the light emitted by an isotope of the element krypton. The meter has since been redefined more precisely as the distance light travels through space in a specified fraction of a second.

A psychological test, as a standard, is even less satisfactory than a meter bar. Since it is a reflection of a culture, its value changes as the culture changes. Further, as its content becomes known to a population of examinees, it produces a defective assessment of the trait it indexes and, in essence, deteriorates.

Any physical device deteriorates with use. Fortunately, a meter stick is only a copy of a standard and when the units wear off, it can readily be replaced by another copy from the master. There are no copies of a psychological test; each test booklet is a master. When a test wears out because of cultural change or test compromise, a new version rather than a new copy must be produced. When this happens, either new interpretations must be made for the new version or the new version must be equated or calibrated to the old version. If comparisons need to be made between old and new test scores, the latter approach must be taken.

The technology of test equating or calibrating has developed because of some urgent needs to make new tests comparable to old tests. To develop a comprehensive solution to the equating problem or even to develop a comprehensive understanding of the problem, it is helpful to explore first these needs for equating. Equating needs are most apparent and equating methods are most extensively applied in the areas of educational and military testing. This review will consider each, in turn.

Equating Needs in Educational Testing

The need for equating in educational settings arises primarily because of the existence of numerous forms of any single test. College admissions tests such as the Scholastic Aptitude Test (SAT), for example, are revised continually to preclude test compromise from one administration to the next. Standardized classroom achievement tests must be revised not only to maintain test security but also to ensure that the content and concepts tapped by such tests are current and relevant to school-district objectives.

Educational Decisions

Decisions concerning individual applications to a university or college are typically made after studying the test scores for all applicants. These scores, obtained from various test administrations, are derived from different forms of an admissions test. Similarly, evaluation of student achievement within one school district requires the accumulation of data across schools and classrooms. Further, questions concerning academic growth and development can be answered only through the administration from grade to grade of test forms whose scores can be interpreted as being from equivalent scales. All these decisions, then, require that meaningful comparisons of scores across test forms be feasible.

Scale Meaningfulness

The raw-score scale of a psychological test rarely has significant implicit meaning. If the raw scores on all forms of a test are expressed in terms of one derived scale (and, therefore, are equated to each other), then the reported scores become independent of the particular test form used to obtain them. There is no requirement that this derived scale have any meaning beyond that of convenience. What is required, however, is that the scale itself remain relatively constant across time.

Consider the history of the SAT as presented by Angoff (1962). The SAT derived scale was originally defined to have a mean of 500 and standard deviation of 100 for the group of applicants taking the SAT in April of 1941. All subsequent SAT forms have been equated back to this 1941 scale.

The number and type of applicants taking the SAT today have changed dramatically since 1941; the scale has long since lost any normative meaning it may have once had. Nevertheless, the constancy of the scale permits comparisons to be made across all SAT forms, old as well as current. In this way, then, colleges can make admissions decisions without special consideration of which test forms were administered.

In a similar fashion, evaluative comparisons based on standardized achievement tests can easily be made across students in a classroom and across schools in a school district. The repeated administration of equated forms also permits inferences to be drawn regarding longitudinal development.

Equating Needs in Military Testing

Equating of military tests has two basic objectives: (a) to make scores on different test forms and composites of test forms comparable and (b) to simultaneously solve the norming problem by relating all scores on new tests and composites back to a wide sample of talent indicative of an anticipated mobilization population. Methodologically, the second problem is subsumed under the first because, if tests can be adequately equated, they can be equated back to the test used on the norming population.

A brief historical overview of military testing may be helpful in explaining the military equating needs. Although military entrance testing extends back to the Army Alpha of World War I, modern testing and equating extends back to the Army General Classification Test (AGCT-1C) and the 1944 mobilization population. Because the United States was then at war, there existed a readily accessible examinee sample representative of the population of draftable personnel that could be tested. The AGCT-1C was administered to a large (approximately 800,000), representative group of military personnel and norms were established on that group.

At that time, each branch of the military had its own entrance examination. While it might be illuminating to follow the development of tests for all of the services, the best documentation exists for those used by the Air Force, and the history of these tests can be reviewed most completely. Weeks, Mullins, and Vitola (1975) reviewed the history of the Air Force entrance examinations from 1948 to 1975, and their review provides an outline for the current discussion.

History of Air Force Testing

Airman Classification Battery

AC-1A. The first operational Air Force recruit classification battery was the Airman Classification Battery (AC-1A), which was implemented in 1948. It consisted of 12 aptitude tests and a biographical inventory. The aptitude tests assessed a variety of characteristics including general aptitudes such as vocabulary and arithmetic skills, general information such as knowledge of current affairs, and vocational skills including knowledge of electronics and mechanics.

Norms for the AC-1A were established by selecting a sample of 1,000 examinees (stratified on the basis of their AGCT composite scores to match the mobilization population) and computing the mean and standard deviation of their scores on each AC-1A subtest. Standard-score stanines were then defined and were used to equate scores on the AC-1A to scores on the AGCT. Additionally, the tests were differentially weighted to form eight composite scores that were used to predict success in military job clusters. These composites were similarly standardized. Thus, the stanine scores were referenced to the 1944 mobilization population.

After the initial standardization, 7 of the 12 aptitude tests were shortened. These, along with the affected composite indices, were restandardized using an equipercentile equating to the original full-length subtests on a sample of 1,018 basic trainees. Thus, the AC-1A, in its final form, had one indirect link in its equating to the mobilization population norms. (The number of indirect links, as used in this report, refers to the number of tests between the equated test and the reference test. A test equated directly to the reference test has no indirect links.)

AC-1B. The AC-1B was a slight modification of the AC-1A. The major changes were the addition of one new test, Pattern Comprehension, and the addition of another composite index, Electronics Technician.

The AC-1A norms were used for all unchanged tests and composites. Norms for the new test and the new composite were obtained by equating the new scores to the AC-1A using the equipercentile method. A composite of two tests was used for the reference in equating the Pattern Comprehension test and another composite was used in equating the new Electronics Technician composite. Thus, the new scores on the AC-1B had two indirect links to the mobilization population norms.

AC-2A. In 1956, the AC-2A was implemented. It consisted of 14 new tests, similar to the previous ones but tapping slightly different aptitudes, and a biographical inventory.

Norms on the AC-2A were obtained by administering the AC-2A along with the AGCT-1C to 2,454 basic trainees (randomly sampled from three Air Force training centers) and equating scores using the equipercentile method. The scoring procedures were also changed for the composites. Composite scores were reported in 20 percentile-based categories rather than nine categories, as before. To accomplish this equating, the AGCT stanines were interpolated and the AGCT scores were considered continuous (Brokaw & Burgess, 1957, p. 11). The test scores themselves continued to be reported in stanines, however. The AC-2A was thus directly linked to the mobilization population.

Airman Qualifying Examinations

The previous Airman Classification Batteries were used primarily for classification rather than selection. When, in 1958, the Air Force implemented a policy of selective recruitment, a new test format was needed. Prior to that time, selection had been done on the basis of the Air Force Qualification Test (AFQT), which was administered at recruiting stations. The basic problem with the classification batteries was that they were too long to administer at the recruiting stations. Thus, a new series of tests was developed, the Airman Qualifying Examinations (AQE). Several versions of the AQE were developed prior to the operational form. (This history is described by Weeks et al., 1975, p. 23.) The form that ultimately replaced the AC-2A was the AQE-D.

AQE-D. The AQE-D consisted of 11 aptitude tests and required just over two hours of testing time, less than half of that required by the AC-2A (Thompson, 1958). These scores were differentially combined into four composite indices. The composites were computed directly from the raw scores on the AQE-D and norms were established only on the four composites. Composite scores were reported in percentiles and were tied to the mobilization population through the AC-2A. The equipercentile method of equating was used, tying each AQE-D composite to the corresponding AC-2A composite. The AQE-D scores thus had one indirect link to the mobilization population.

AQE-F. The AQE-F replaced the AQE-D in 1960. The battery content remained the same except for the substitution of a Hidden Figures test for the Figure Recognition test. Like the AQE-D, the composite indices were equated to the AC-2A using the equipercentile method. The AQE-F composite scores thus had one indirect link to the mobilization norms.

AQE-62. The AQE-62 replaced the AQE-F in 1962. The major content change involved replacing the Clerical Matching and Numerical Operations subtests with an arithmetic test. Thus, the AQE-62 contained ten subtests. It was normed through equipercentile equating using the AQE-F, on a group of 2,428 basic trainees, as the reference test. Again, the composite rather than the subtest scores were equated (Edwards & Hahn, 1962). The AQE-62 thus had two indirect links to the mobilization population norms.

AQE-64. In 1960 Project TALENT, which was sponsored by several Government agencies, provided a new reference population to use for norming tests. In the spring of that year, a comprehensive battery of aptitude, achievement, background, interest, and personality tests was administered to a sample of more than 400,000 high-school students. The test battery contained 74 tests and had 82 scores. Since there was reason to believe that the 1944 group was no longer appropriate as a reference population, the subgroup of the TALENT sample consisting of high-school seniors was chosen as a new reference group comparable to the mobilization population in intellectual abilities and educational attainment.

For AQE equating, TALENT test composites were developed to predict each of the four Air Force composites (Dailey, Shaycoft, & Orr, 1962). This was done using stepwise multiple regression in a sample of 2,489 basic airmen. The sample was divided into two subsamples of nearly equal size and separate regressions were run. Three to four TALENT tests were chosen to predict each of the AQE composites. With a few exceptions, the first tests stepped in were chosen for the most predictive composites. Some non-statistical considerations also weighed into the selection.

The AQE-64, which replaced the AQE-62 in 1964, was similar to the AQE-62. It differed in that its unspeeded Arithmetic test was replaced with a speeded Arithmetic Computation test and the composites were revised to include educational variables (Madden & Lecznar, 1965). It was normed relative to the TALENT sample using each of the four AQE composite indices in four groups of approximately 1,000 basic trainees each. An equipercentile equating was then done between each AQE composite index and the corresponding TALENT index. The AQE-64 thus had a direct link to the new norm group and no link to the 1944 mobilization population.

AQE-66. The AQE-66, which replaced the AQE-64 in 1966, was essentially identical to the AQE-64. It contained new items and the Arithmetic Computation test's content and its order in the administration sequence were changed. No substantial changes occurred. It was equated to the TALENT tests using four groups of approximately 1,000 basic trainees each and the same equipercentile procedures that were used on the AQE-64.

AQE-J. The AQE-66 was replaced in 1971 by the AQE-J. No changes, other than the items, were made in this revision. It was again normed by equating it to the TALENT battery on four samples of aproximately 1,000 basic trainees each.

Armed Services Vocational Aptitude Battery

All Armed Services entrance batteries were replaced in 1973 with the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB was similar in coverage to the AQE. The ASVAB-3, the version initially implemented for operational testing, consisted of nine tests covering basic aptitudes with tests like Coding Speed and Word Knowledge, and measuring vocational achievement with tests like Shop Information and Electronics Information.

From the ASVAB, the Air Force computed the same four composites that it computed from the AQEs. These composites were equated to the TALENT composites and thus to the TALENT norm group. This was accomplished by administering, in four groups of approximately 1,000 basic airmen, the ASVAB and the TALENT tests required to calculate one composite score (Vitola & Alley, 1968). Each ASVAB-3 composite was equated in its respective group of examinees to the appropriate TALENT composite using the equipercentile method.

In more recent years, the Air Force shifted back to calibration against the AFQT. Ree, Mathews, Mullins, and Massey (1982) described an equating study in which Forms 8, 9, and 10 of the ASVAB, versions A and B, were equated back to the AFQT-7A. The AFQT composite of the ASVAB was the score equated. The single-group procedure was used along with the equipercentile transformation. Because of a lack of appropriate literature, it was not possible to determine the norm group and linkage techniques used for the AFQT-7A.

Military vs. Educational Testing

The history of Air Force testing suggests several differences between equating in a military setting and in an educational setting. First, it is primarily composites that are equated in the military, rather than individual test scores. Although batteries usually include eight to ten subtests, the history of the Airman Classification Batteries shows that it has always been the four composite indices that have been equated. This differs from the educational environment, where subtests rather than composites are usually equated.

Second, the military relies exclusively, it appears, on the single-group data collection design in which a single group of individuals takes both of the tests to be equated. This differs strikingly from the educational environment, which relies heavily on anchor-test methods, and is probably due to the military's captive group of examinees. The ability to assemble a large group of examinees for single-group equating may be a great advantage in developing a superior equating method.

OVERVIEW OF EQUATING MODELS

Calibration vs. Equating: A Clarification of Terminology

The literature contains some dispute and confusion regarding what can properly be called equating. Probably the most comprehensive discussion of classical equating procedures was provided by Angoff (1971), who considered both the conceptual problems and some practical designs for equating. Angoff discussed the general problem of making scores comparable and divided this into the two processes: calibration and equating. Angoff took the position that for tests to be equated they must be equivalent. That is, they must measure the same characteristic with the same reliability and be of equal difficulty. Essentially, they must be parallel. For all other cases, he preferred to use the term calibration, which referred to putting scores on a common scale without calling them equivalent. This is consistent with Lord's (1977, 1980) definition that two tests can be equated only if it can be considered a matter of indifference to an examinee which test he or she takes.

Calibration

Calibration, in Angoff's discussion, referred to a procedure for putting different measurements on a common scale. He used, as an example, the physical dimension of temperature with several different thermometers corresponding to tests. If a fever thermometer, a refrigerator thermometer, and an oven thermometer are considered, they obviously are not interchangeable. Values from one cannot be equated to values of another. Using standard equipercentile procedures, all scores on the refrigerator thermometer would map into the lower bound of the fever thermometer. All these thermometers can be calibrated to a standard temperature scale, however.

Implicit in this example is the fact that equating is not necessarily a more desirable procedure than calibration. Most psychometricians would be very happy with a set of tests that scaled like thermometers, even if they could not be equated to each other.

Equating

Lord (1980) and Angoff (1971) agreed that there are three requirements for equating tests that measure the same ability: (a) symmetry, (b) invariance, and (c) equity (cf. Cook & Eignor, 1983). The requirement of symmetry implies that the equating transformation should be the same regardless of which test is labeled *x* and which test is labeled *y*. Hence, all regression methods are inappropriate for test equating since, in general, the regression of *x* on *y* differs from the regression of *y* on *x*.

The property of invariance requires the equating transformation to be unique regardless of which population subgroup was used to derive it. Angoff (1971) has shown that this is, in general, true for (fallible) observed scores only if the two tests are strictly parallel.

Lord's (1977, 1980) definition of equity requires that it be a matter of indifference to an examinee which test he or she takes. This implies that the tests must be equally reliable. Modern test theory (Lord, 1980; Lord & Novick, 1968) suggests that tests are not equally reliable across the ability range but tend to be more reliable at the ability levels appropriate for their difficulty. Thus, two tests of unequal difficulty cannot be equally reliable at all ability levels and thus cannot be equated.

Test equating, then, under the strict definitions of Lord and Angoff, appears to be of little practical value. For tests to be equated they must be perfectly reliable or strictly parallel. If tests are strictly parallel, however, they will not need to be equated.

It may be that the concept of equating, and thus the term, is too limiting for practical problems and that calibration would be a more appropriate term. On the other hand, the term equating has been applied to a wide range of comparability efforts and even though calibration may be a more appropriate term, it is unlikely that the psychometric community will readily accept the change. It thus appears more profitable to revise the definition of equating to one that is more in line with the practical goals of the procedures.

Operational Definition of Equating

Theories of psychological traits imply an underlying dimension of a characteristic that tests attempt to assess. The concept that an individual has a level of the trait is implicit in the "true score" of classical test theory. The true score is not a procedure-free measure of the trait level, however, because it relates to a specific test. Modern test theories allow a procedure-free trait value to exist, at least conceptually. Since in concept it is the assessment of this trait level that is the objective of testing, it seems reasonable to say that two tests are equivalent if two equated scores result from the same trait level on both tests. That is, two tests are equated if each trait level leads to equivalent scores on the two tests. This definition would be considered a form of calibration by Angoff. It is more in line with the goal of what is commonly called equating, however, and will be used as a definition of equating throughout this report.

The term trait, as used above, should not be considered limited to a unidimensional trait or even to a single trait. The concept is general and, in the case of several traits leading to a composite, can be considered to mean that, for any fixed set of trait levels, two composites are equated if the expected values of their scores are equal.

Data Collection Designs

Methods for equating scores can be classified on the basis of two factors: the design by which data are collected and the methods by which the equating transformation is determined.

Angoff (1971) listed six major equating designs. In terms of data collection, these six designs can be grouped into two categories: designs assuming equivalent samples of examinees to achieve equation (Designs I and II) and designs employing an anchor test to achieve equation (Designs III, IV, V, and VI). Design I assumes that each test was given to one of two random samples from a population. In Design II, both tests are administered to a single random sample, thus resulting in a single-group design. Test forms are counterbalanced during admininstration to prevent order effects. Design III is really a combination of an equivalent-groups and an anchor-test design. Two random groups are selected from a population and one test and the anchor test are administered to each group, as in Design 1. The anchor test is used to estimate test-score statistics for the combined group of examinees. Design IV is similar to Design III; the difference is that the groups are not random samples from a population. In Design V, each test is equated to a common anchor test; scores that are equated to the same anchor test score are considered to be equated to each other. Design VI is similar to Design IV in terms of data collection; here different scaling methods are applied to common (anchor) items.

Marco (1977) and Marco, Petersen, and Stewart (1980) listed three data collection designs: (a) all items are given to a single group of examinees; (b) the same set of items is administered to different groups of examinees; and (c) an anchor set of items, either internal or external to the tests to be equated, is administered along with all tests given to different groups of examinees.

A recent study of item response theory (IRT) parameter linking (Vale, Maurelli, Gialluca, Weiss, & Ree, 1981) suggested four basic data collection designs of potential utility for equating: (a) the equivalent-groups method, (b) the equivalent-tests method, (c) the anchor-group method, and (d) the anchor-test method. In the equivalent-groups method, a population of examinees is randomly split into two or more groups and each group is given a different test.

Equating is achieved by assuming that the groups are equivalent and adjusting the test scores such that the score distributions of the two tests are identical. In the equivalent-tests method, a domain of items is randomly split into two or more tests, and these tests are given to different groups of examinees. The tests are assumed to be randomly equivalent, and no explicit equating is done (i.e., it is assumed that the scores are already equated). The anchor-group method employs a common group of individuals to take all tests to be equated. Equating is done using this group in the same manner as with the equivalent-groups procedure. A common set of items is used by the anchor-test method. Equating is accomplished by equating the non-anchor tests to the anchor test in what amounts to several single-group procedures. Angoff's first two designs are examples of the equivalent-groups method, and his latter four are examples of the anchor-test method. Marco's first design is a special case of the equivalent-groups method, his second is a special case of the equivalent-tests procedure, and his third design is an application of the anchor-test method.

Vale et al. (1981) suggested that the equivalent-groups and anchor-test methods were most useful in linking applications. Linking, the mapping of item parameters onto a common scale, differs from equating in several ways. Most importantly, linking is properly implemented only using IRT procedures and is typically applied to sets of items of similar difficulty and examinee groups of similar ability. For reasons different than those cited by Vale et al., the same two methods are probably most useful for equating. The equivalent-tests procedure is inappropriate for equating because it amounts to assuming the conclusion (i.e., that the scores on the tests are equivalent). The anchor-group method, while applicable to IRT linking procedures where linking is separate from item calibration, would be indistinguishable from a single-group procedure in an equating design; this has already been included as a special case of the equivalent-groups design.

Equating Transformations

The transformation in equating is the function that maps scores from one test onto the other test. Several different transformations have been proposed.

Conventional Equating

Both linear and equipercentile methods are used to equate conventionally administered and scored tests.

Equipercentile Equating

Raw transformation. A procedural definition of equating (as distinguished from the conceptual ones discussed earlier), provided by Lord (1950) and Flanagan (1951) and reproduced by Angoff (1971), states that "Two scores, one on Form X and the other on Form Y (where X and Y measure the same function with the same degree of reliability), may be considered equivalent if their corresponding percentile ranks in any given group are equal" (Angoff, 1971, p. 563). This procedural definition gave rise to the equipercentile transformation which is accomplished by assigning an equal value to scores on two tests when the same percentage of individuals falls below these scores (i.e., such that equated scores have equal percentile ranks). Procedurally, this transformation is typically performed on observed scores.

Angoff's (1971) Design I for equipercentile equating can be used to equate conventionally scored parallel tests using both the single-group and equivalent-groups data collection designs. Angoff's Design V can be used for the anchor-test design; it equates the old and new tests separately to the common anchor test using the single-group equipercentile method, and then defines as equated those scores on the old and new tests equivalent to the same anchor-test score.

Smoothing. Observed cumulative percentiles may exhibit irregularities because of sampling and measurement error; smoothing the data may yield better equating results. There is no systematic evidence to indicate which smoothing method is optimal or when in the equating process smoothing is best applied. That is, the individual frequency or percentile tables may first be smoothed, with these smoothed tables then used to equate the two tests. Alternatively, the unsmoothed tables can be used to equate the tests; the resulting equating transformation itself can then be smoothed.

The analytic methods of smoothing include cubic polynomial regression and cubic spline functions. In regression smoothing, either the raw percentiles or equated scores are regressed (using cubic polynomials) on observed test scores; inserting these test-score values back into the resulting regression equation yields either a smoothed frequency curve or a smoothed equating transformation, respectively.

Cubic splines are part of a family of functions used to fit curves to observed data. In cubic-spline smoothing (Reinsch, 1967), a separate cubic spline function is fit to each interval between adjacent score points. For score points x_i, where $\underline{i}$ ranges from 0 to the maximum test score, $\underline{k}$, the general form of the spline function over the interval $x_i < x < x_{i+1}$ can be expressed as

$$f_i(x) = a_{0i} + a_{1i}x + a_{2i}x^2 + a_{3i}x^3 \qquad [1]$$

where the constants a_{0i}, a_{1i}, a_{2i}, and a_{3i} can vary from interval to interval and the spline functions meet at their common endpoints. The spline functions are constrained such that the second derivatives ($f''_i(x)$) of the (two) spline functions at each score point are identical. The spline function over all x values is denoted f(x); spline functions minimize

$$\int_{x_0}^{x_k} |f''(x)|^2 dx. \qquad [2]$$

The degree of smoothing can be explicitly controlled by the user and is dictated by the value of a smoothing parameter, S; S controls the degree to which the spline function values are permitted to deviate from the observed values (y_i). The smoothing parameter is directly proportional to the differences between the observed and smoothed values and is inversely proportional to the relative weights assigned to the score points ($\delta(y_i)$). The differences between the observed and smoothed values at each score point can be unit weighted or, alternatively, weighted by the standard errors of the scores. All spline functions, then, minimize Equation 2 subject to the restriction

$$\sum_{i=0}^{k} \left(\frac{f(x_i) - y_i}{\delta(y_i)} \right)^2 \leq S. \qquad [3]$$

The choice of values for S and the score-point weights determines the specific nature of the final smoothing solution. The larger the value chosen for S, the greater the degree of smoothing; if S is set equal to zero, the smoothed values equal the observed values and this process becomes a cubic-spline interpolation method. Reinsch suggested using standard-error weights (if they are available) for the score points and a value for S within the range $K \pm (2K)^{1/2}$ where $K = k + 1$ is the number of score values. Kolen (1983) argued that the observed percentiles/equated score points are not independent in the context of test equating and, therefore, that Reinsch's suggested S values may be inappropriate. He suggested, instead, a "moderate" smoothing parameter that is equal to one-half the number of score points.

There are problems inherent in applying either type of smoothing method to examinee test data. The cubic-spline method depends heavily on the choice of a value for the smoothing parameter; at this point, there are no standards for selecting this value. Applying either type of smoothing method to the equating transformation itself yields a nonsymmetric equating: The resulting smoothed transformation equating Test X to Test Y is not the same as the smoothed transformation equating Y to X. However, the goal of smoothing is to eliminate error-induced irregularities and discontinuities in the observed data; because lack of symmetry is not a concern at that point, it may make the most sense, theoretically and practically, to smooth the raw frequency or percentile tables rather than the equating transformation itself (as has been the usual practice).

Linear Equating

The linear equating method has typically been used as an approximation to the equipercentile method for equating observed test scores. In the linear method, scores on two tests are considered equivalent if they correspond to the same standard (i.e., *z*) score (Angoff, 1971). This is equivalent to the equipercentile procedure only if the distributions of test scores on the two tests are identical. Practically, it often makes a good approximation to the equipercentile result if the distributions are similar in shape. Furthermore, the linear method is less sensitive to sampling fluctuations with extreme scores that occur in the tails of distributions.

Problems with Conventional Equating Methods

There are two aspects of Angoff's (1971) conventional definition of equating that make these transformations only approximately correct. The first is "any given group," which requires (a) that further assumptions concerning the test scores be made before equating, or (b) that an exhaustive sampling of all possible groups be made. The second aspect is the requirement of equal reliability, which implies that tests which differ in reliability cannot be equated so simply.

The any-given-group restriction is satisfied if the relative shapes of the score distributions remain constant across groups at all levels of the trait. According to Lord (1977), it is a necessary (but not sufficient) condition for accurate test equating that the percentile ranks of equated scores remain equivalent across all groups tested.

The problem of equating unequally reliable tests has been addressed in several ways. Since the problem of unequal reliability disappears when true scores are equated, methods of transforming true scores have typically been applied. Angoff (1971) provided linear procedures for equating unequally reliable tests. In general, the procedures he suggested replaced the standard deviation in the standard-score formula with the standard deviation multiplied by the square root of the reliability coefficient (cf. Angoff, 1971, p. 571).

Classical test theory does not provide a mechanism for adjusting for reliability in equipercentile equating. The true-score regression suggested by Angoff could be applied prior to equipercentile equating but would result in exactly the same equating transformation as if it were not applied at all. For example, a z score of 1.0 might correspond to a percentile rank of 84. If the reliability were 0.81, the true score would be regressed to 0.90. Its percentile rank would still be 84, however, and since the equipercentile correspondence is between percentile ranks, it would still correspond to the same score on the second test.

Item Response Theory

Item response theory (Birnbaum, 1968; Lord & Novick, 1968) offers another method of equating tests that differ in difficulty and reliability. IRT expresses the probability of a keyed response as a function of an examinee's trait level and one or more characteristics of the item. An IRT model often used with dichotomous items is the three-parameter logistic model. This model describes the probability of a correct response to an item as a function of the trait level (θ), the item's discriminating power (a), its difficulty (b), and its proneness to being answered correctly through guessing (c):

$$P(\theta) = c + \frac{1-c}{1 + \exp(-1.7a(\theta-b))} \qquad [4]$$

The one-parameter (or Rasch) IRT model describes the probability of a correct response solely as a function of theta and item difficulty. The two-parameter IRT model includes theta and the a and b parameters only. No provision is made in these models for answering an item correctly through guessing. When all of the parameters of an IRT model are estimated, either true-score or observed-score IRT equating can be performed.

True-Score Equating

Once the item parameters and theta for two or more tests are expressed on a common metric, the relationship between ability (θ) and number-correct true score (ξ) on Test X can be expressed as:

$$\xi = \sum_{i=1}^{n_x} P_i(\theta) \qquad [5]$$

Similarly, the relationship between ability and number-correct true score on Test Y (η) can be expressed as:

$$\eta = \sum_{j=1}^{n_y} P_j(\theta) \qquad [6]$$

For any given value of θ, the corresponding true scores ξ and η are equated in an exact (mathematical, not statistical) way (Lord, 1977). The equating transformation relating ξ and η is typically obtained by estimating all relevant item parameters (e.g., by obtaining $P_i(\theta)$ for all $\underline{i}$).

Because this equating transformation is derived from true number-correct scores, it should, strictly speaking, be applied only to examinee true scores. However, only observed scores are available to test researchers and administrators. The typical procedure is to estimate the true number-correct score for each examinee and apply the IRT true-score equating transformation to these score estimates. While this is theoretically inappropriate, its utility remains an empirical question.

Observed-Score Equating

Alternatively, one may estimate the frequency distribution of number-correct scores for each test for the combined examinee group by using (a) the estimated distribution of ability (theta) in the combined group, (b) estimated IRT item parameters for each test, and (c) the generalized-binomial generating function for obtaining the frequency distribution of scores conditional on theta (cf. Lord, 1977, p. 131). Once the frequency distribution of scores on each test has been estimated for the combined examinee group, the observed test scores can then be equated using ordinary equipercentile procedures.

Item Parameter Estimates

The parameters of IRT must, of course, be estimated. Practically, calibration programs typically assume theta to be distributed with a mean of zero and a standard deviation of one. In order to equate two tests by the methods described above, the parameters of two tests calibrated separately must be linked together onto a common metric. The problem of making the score metrics equivalent has thus simply been shifted from the true scores to the thetas. The parameter-estimation problem still exists, although it is somewhat less severe than in the strong true-score methods. The main advantage of IRT, in this respect, is that the theoretically proper linking transformations for the theta metric are linear and can be made more accurately. A major disadvantage of IRT equating is that it assumes the trait to be unidimensional. Methods of linking item parameters have been discussed in detail elsewhere (Vale et al., 1981).

Strong True-Score Theory

Lord (1965, 1969, 1980) developed a strong true-score theory (STST) to produce an estimated distribution of true scores ($g(\zeta)$) from a distribution of observed scores ($\phi(x)$). Tests are equated by applying conventional equipercentile procedures to the true-score distributions. The general model is expressed in terms of the equation

$$\phi(x) = \int_a^b g(\zeta) h(x|\zeta) d\zeta \qquad [7]$$

where $h(x|\zeta)$ is the conditional distribution of observed scores on true scores. The limits of integration, a and b, are set at the practical true-score limits, 0 and 1, for a proportion-correct true score.

Strong true-score theory attempts to describe the distribution of true scores, $g(\zeta)$, by solving the integral equation with specified functions for $\phi(x)$ and $h(x|\zeta)$. For dichotomous test items, Lord has typically used a compound binomial, or an approximation to it, for $h(x|\zeta)$. Solving the integral equation is a numerically tedious procedure and Lord has taken two approaches to it. In the first (Lord, 1965), instead of trying to define an entire nonparametric distribution of true scores, he assumed an incomplete Beta distribution for $g(\zeta)$ and estimated its parameters. This was numerically simpler than a nonparametric approach. It appeared to work well in cases where the a and b estimates from the observed-score data fell within the appropriate limits of 0 and 1, but did not function well otherwise.

Lord (1969) attempted a more empirical approach and tried to develop for $\phi(x)$ a polynomial of degree as high as was warranted by the data. He found that this procedure worked well when the test was administered to at least 10,000 examinees. Such large samples are, unfortunately, often unavailable.

In principle, if the true-score distribution functions of two tests could be estimated by these methods, the problems of test unreliability would disappear and test equation could be achieved by setting the true-score distribution functions equal to each other:

$$\int_0^x g(\xi) d\xi = \int_0^y g^*(\zeta) d\zeta \qquad [8]$$

The values x and y that solved the equation would thus be equivalent scores.

The strong true-score theory approach to equating is, theoretically, a general solution to the problem of equating nonparallel tests. Problems resulting from reliability and difficulty differences disappear when true scores are used. The assumption of a compound binomial conditional distribution of observed scores given true scores does not seem to be overly restrictive. The numerical problems encountered in solving the equations are formidable, however, and the statistical estimation procedures available are unsatisfactory. First, Lord used a numerical approximation to the compound binomial distribution because exact evaluation required too much computation. Then, numerical procedures were required to solve the integral equation. Finally, numerical procedures were required to equate the integrals. The statistical procedures require such a large number of examinees for Lord's (1969) empirical procedure that it is, typically, impractical. These problems may not be insurmountable but suggest, at a minimum, that the strong true-score theory procedures are considerably more difficult to apply than are the conventional procedures.

EVALUATION OF EQUATING METHODS: A REVIEW

The published literature describing applications of equating methods provides some information regarding the relative utility of the various equating methods and helps to identify the most frequently reported equating problems. Furthermore, these past studies help to identify the methodological issues and problems crucial to the design, execution, and evaluation of the current research. These issues include, but are not limited to, (a) characteristics of the ability and score distributions in the population studied, (b) the psychometric characteristics of the tests, (c) procedures for combining tests into composites, and (d) definitions of evaluative criteria.

This review of equating applications is divided into several sections. In the first, conventional applications (linear and equipercentile methods) are described and critiqued. IRT equating efforts are discussed in the second section. The third section considers studies that compared conventional and IRT methods. The remaining sections summarize the findings from previous research and discuss their relevance to practical equating situations. The criterion problem is discussed. No studies investigating STST methods were found; STST is thus not included in this review.

Previous Research

Conventional Equating Methods

Regression

If two tests are strictly parallel, their true-score distributions will be identical, and the regression of Form X true scores on Form Y observed scores will be identical to the regression of Form Y true scores on Form X observed scores. Hence, the observed scores from a single test form can be used to equate the true-score distributions of all parallel forms. However, the true scores of nominally (i.e., imperfectly or nearly) parallel tests (Lord & Novick, 1968, p. 174) may be highly correlated but are not identically distributed, and the above relationships do not hold.

Marks and Lindsay (1972) conducted a Monte Carlo simulation to investigate the effects of violating the strict-parallelism assumption on equating adequacy. They varied test and sample characteristics and examined the accuracy of estimating the true-score distributions of one test form from the observed scores of a second form. They manipulated sample size (100, 250, 500), test length (30, 60, 120), test-form reliability (.75, .85, .95), and the correlation between true scores (.80, .90, 1.0); these four parameters were completely crossed in their study. Observed scores on each test form were computed for each simulated examinee using (a) bivariate normal true-score distributions with the specified correlation, (b) appropriate error-score distributions, and (c) the classical test model $\underline{X} = \underline{T} + \underline{E}$, where $\underline{X}$, $\underline{T}$, and $\underline{E}$ represent observed, true, and error scores, respectively. True scores were then estimated on Form X as might be done in a practical setting, using the equation

$$\hat{T}(X) = \rho_{XX'}X + (1 - \rho_{XX'})\,\overline{X} \qquad [9]$$

where $\rho_{XX'}$ is the reliability, and $\overline{X}$ is the sample mean. The true scores of Form Y were also estimated according to Equation 9. Equating was accomplished by regressing the estimated true scores of Form X onto the estimated true scores of Form Y.

Marks and Lindsay performed a four-way analysis of variance (pooling the 3- and 4-way interaction terms) on the data, using as a dependent variable the mean squared difference between the estimated true score on the equated test (i.e., after the regression) and the actual true score generated in the simulation. They concluded that sample size was the most important factor affecting equating error; as sample size increased, it "washed out" the effects of the other test-form characteristics. They discouraged the use of sample sizes smaller than 250 when equating nominally parallel tests. The effect of test-form reliability was not statistically significant.

Because Monte Carlo methods were used in this study, true scores on each test form were known for each examinee. Thus, the equating procedure could be readily evaluated by comparing the estimated (i.e., equated) true score to the specified true score for each examinee. Such a criterion does not exist when test scores from real examinees are equated.

This study is seriously flawed in other ways, however. Regression methods, as used by Marks and Lindsay, are not appropriate for equating two tests because they violate a basic equating requirement. That is, equated tests must be symmetrically related, and the results of an equating procedure should be the same regardless of which test form is labeled X and which test form is labeled Y. However, the regression equation that predicts the score on Form X from the score on Form Y is not the same as the regression equation that predicts Y from X (cf. Lord, 1980, pp. 198-199). Therefore, the equating results obtained by Marks and Lindsay would be different had they regressed Y onto X. This is an unsatisfactory consequence of the regression procedure.

Linear

Garcia-Quintana and Johnson (1979) compared three methods of linear equating designed for use with parallel tests administered to nonequivalent examinee groups along with a common anchor test. One of two forms of the SRA Mastery Mathematics Tests was administered to more than 2,000 sixth-grade students; all students received the Mathematics Test of the Comprehensive Tests of Basic Skills (CTBS).

Their first linear equating method involved procedures from Angoff's Design IV (using equations attributed to L. R. Tucker; cf. Angoff, 1971, p. 580), where the CTBS anchor-test scores were used to adjust for ability differences between the groups before equating the standard scores on the two SRA test forms. Design IV requires that summary statistics for each test form for both groups combined be first estimated from the combined distribution of anchor-test scores before the linear transformation is applied. The other two methods used procedures from Angoff's Design V to equate both test forms to the anchor test and to define as equivalent (a) the scores on the two tests that were equated to the same anchor-test score and (b) the scores on the two tests that were predicted by the same anchor-test score.

The equating tables derived from these three methods were compared with each other for consistency because there was no external criterion of equating adequacy. The authors found that these methods yielded similar results throughout the middle score range, with differences among the three methods becoming more pronounced at the extremes of the score distributions.

Equipercentile

Yen. Yen (1982) applied the equipercentile equating method to data generated according to the three-parameter logistic IRT model. Test length (n), sample size (N), and differences in item difficulties and discriminations were varied across simulated tests. These factors were completely crossed. Each sample of size N was generated so that true theta was distributed standard normal. For each pair of tau-equivalent tests (having equal expected means) in a given condition, N pairs of theta estimates were generated; one set for each test. Each theta estimate was chosen by using a normal random deviate generator, assuming theta estimates for a given test and theta value to be normally distributed with mean equal to true theta and variance defined by the inverse of the information value (calculated using the appropriate item parameters and true theta). Equipercentile equating was then performed for each test pair.

Yen used a bias measure as a criterion, defining

$$\hat{\beta}_i(X_1, X_2 | \theta) = \frac{\bar{X}_{1i} - \bar{X}_{2i}}{s_{X_i}} \qquad [10]$$

where X_1 is the raw score on the first test, X_2 is the equated score on the second test, and S_{X_1} is the standard deviation of X_1 computed over all examinees. Simulated examinees were divided into five cells by rank order on theta, and two summary indices were considered: maximum absolute bias and mean absolute bias. Equipercentile-equating bias was computed from the difference between the theta estimate from the first test and the corresponding equipercentile-equated theta estimate from the second test; bias between the two sets of paired (tau-equivalent) theta estimates was computed as a basis for comparison.

Errorless equating of the theta estimates, according to the IRT model, would have resulted in a linear transformation function. Curvilinearity of the equipercentile plots was quantified by subtracting the Pearson correlation coefficient from the average of the two correlation ratios (eta). Curvilinearity increased (and therefore goodness of equating decreased) as the length of the test decreased and as the mean difference between test difficulties increased. For tests of equal difficulty, the equipercentile bias was less than the IRT "bias," but bias was substantial otherwise and increased as test length decreased and the disparity between the test difficulties increased. Differences in item discriminations across tests did not adversely affect equipercentile equating, at least for the high levels and small differences simulated here. Test-length differences (20 vs. 40) were important while sample-size differences (1,000 vs. 2,000) were not. No comparisons were made with any other equating method.

Slinde-Linn. The Anchor Test Study (Bianchini & Loret, 1974) was a large-scale study designed, in part, to equate seven standardized reading tests to each other within three grade levels (4, 5, 6); an eighth test was added later. It did not, however, equate the tests across grade levels. Slinde and Linn (1977) used data from the Anchor Test Study and test publishers' norms to investigate the adequacy of equipercentile equating methods and the anchor-test data collection design for vertical equating situations (i.e., where tests differ widely in difficulty and examinees differ widely in ability). Because the standardized tests from different publishers changed forms at different grade levels, a variety of equating comparisons were possible.

In each case, different levels of the same test were equated using various other published tests as anchors. In all cases, differences between scores on a single level of a test were not the same as differences between scores on vertically equated levels of the same test. The direction of this difference was not consistent across tests and levels. Slinde and Linn acknowledged that some of their results may have been confounded because publishers' norms rather than

Anchor Test Study norms had to be used at times to define the scaled scores. Nevertheless, they suggested that other equating methods (e.g., IRT) might be better suited to the task of vertical equating.

Comparisons Among Conventional Methods

Bianchini-Loret. The original Anchor Test Study (Bianchini & Loret, 1974; see also Linn, 1975) was a monumental endeavor designed to equate scores across eight widely used standardized tests and to provide new national norms for each of those tests. It also allowed for a comparison of different equating procedures. To this end, pairs of tests were administered to different groups of examinees, and equipercentile equating methods were compared to linear methods. The full sample and eight balanced half-samples were used to equate each test to one of the other reading tests; the root mean squared deviation of the equivalent scores for each half-sample replication from the equivalent scores was computed for the full sample. According to this error-of-equating criterion, the equipercentile methods were found to be superior, with an estimated equating error generally less than one raw-score point (except in the chance-test-score range).

Stock-Kagan-Van Wagenen. Stock, Kagan, and Van Wagenen (1980) equated verbal, quantitative, and composite scores on the Graduate Record Examination (GRE-V, -Q, and -C, respectively) to scores on the Miller Analogies Test (MAT) from the responses of 273 graduate-school applicants who took both tests. Four different equating procedures were employed and compared: (a) MAT scores were regressed on GRE-V, GRE-Q, and GRE-C separately, and vice versa; (b) conditional mean scores on the three GRE subtests were obtained for each MAT score, as was the mean MAT score for each GRE subtest score (i.e., a form of curvilinear regression was performed); (c) linear equating was performed between the MAT and each GRE score; and (d) equipercentile equating was performed between the MAT and each GRE score. The equating tables derived from these four procedures were examined and compared with each other.

Stock et al. observed several deviations from monotonicity in equated scores using conditional means. Although this was probably due to sampling fluctuation and should therefore disappear with larger samples, they dismissed the method from further consideration. As expected, scores equated with the regression procedure were closer to the mean than were scores equated with any of the other procedures. Linear and equipercentile methods yielded virtually identical results, with the simpler linear method therefore preferred by the authors. Stock et al. concluded that the ultimate choice was between the linear and regression procedures, with regression preferred if the correlation between the two sets of scores was available (i.e., for the single-group data collection design). This recommendation was made in the absence of any external criterion on which to compare the methods.

As discussed previously, regression methods (linear or otherwise) are inappropriate for equating two tests since they do not yield a symmetric equating transformation. Hence, the only valid comparison in this study is the one between the linear and equipercentile procedures. Strictly speaking, neither of these methods was appropriate for equating in this situation. These procedures assume that the two test forms are parallel. This obviously was not the case for the GRE and MAT. The content of the MAT differs greatly from that of the GRE. The correlations of the MAT with the GRE scores ranged from a low of .42 (with GRE-Q) to a high of only .70 (with GRE-V). In any case, because of the lack of any criteria for evaluation, conclusions regarding the relative merits of these two equating procedures cannot be drawn.

Lord. Lord developed formulas for the standard errors of equating for tests linearly equated using an anchor test (1975) and tests equated by the equipercentile method (no smoothing) using a single-group or equivalent-groups design (1981a, 1982c). He considered these indices to be on the same scale and used them to compare equipercentile to linear equating (Lord, 1981a, 1982c). Scores on two forms of the SAT-V were equated by both methods using an external 40-item anchor test; each test form was administered to nearly 2,700 examinees.

Lord used Angoff's (1971) Design IV to linearly equate the SAT-V forms. Each form was also equated to the anchor test using equipercentile procedures; scores equated to the same anchor-test score were assumed to be equated to each other. This method of equipercentile equating is, essentially, two applications of single-group equating. Consequently, the (independent) sampling variances as defined in Lord (1981a, 1982c) were summed together.

The standard errors computed for the different equating methods were then studied. For both the linear and equipercentile methods, standard errors were smaller for scores in the middle of the range. Standard errors of the equipercentile equating were approximately twice as large as those of the linear equating for middle-range scores; this difference became even larger in the tails of the score distribution.

Lord also compared the equating tables resulting from application of the two equating methods. The two sets of equated scores were similar in the middle of the score range but were quite disparate (more than one standard error of measurement apart) in the tails. This difference corresponded to ten standard errors for the linear equating and nearly five standard errors for the equipercentile equating at that score level.

Summary

The results of these studies suggest that there are few, if any, practical differences among the conventional equating procedures. Regression procedures are clearly inapppropriate for test equating (cf. Marks & Lindsay, 1972; Stock et al., 1980), but this result should be obvious without empirical study. Equipercentile and various modifications of linear methods generally yield similar results (Garcia-Quintana & Johnson, 1979; Stock et al., 1980), although Bianchini and Loret (1974) found the equipercentile methods to be superior in terms of cross-sample replication. Lord (1981a, 1982c) observed larger standard errors for equipercentile equating in his study, although it is not known whether the same results would have been obtained with equating lines that had been smoothed in some way.

It is questionable whether equipercentile procedures can be successfully used for vertical equating -- that is, for equating tests that differ in difficulty (Slinde & Linn, 1977; Yen, 1982). Linear procedures were not used to vertically equate tests in the studies reviewed here.

IRT Equating Methods

One-Parameter Model

Slinde-Linn-Gustafsson. Slinde and Linn (1978, 1979a) presented a set of studies designed to evaluate the adequacy of the one-parameter IRT model for vertical equating. In their 1978 study Slinde and Linn used response data from 1,365 examinees on a 36-item mathematics achievement test. Two tests of differing difficulty were obtained by dividing the 36-item test into two 18-item tests on the basis of the item difficulties obtained in the group of 1,365 examinees. The average proportions correct for the tests were .665 for the easy test and .362 for the difficult test. The examinees were then divided into low-, middle-, and high-ability groups on the basis of their scores on the easy test.

Item difficulty parameters were calculated for the total set of 36 items in the low-ability group, the high-ability group, and the total group (the middle-ability group was reserved for later use). Ability estimates were then calculated for each of these groups (low, high, and total) using parameters obtained from each group in a crossed design. Mean differences between ability estimates derived from the easy test and the difficult test were then computed and compared.

When the total-group ability estimates were calculated using item parameters obtained from the total group, the difference between means obtained from the easy and difficult tests was trivial. Similarly, when the high-group mean ability estimates were calculated using item parameters obtained from the high group and when the low-group means were calculated using the item parameters obtained from the low group, the differences were trivial. When items calibrated in the high group were used to estimate abilities in the low group or the middle group and when items calibrated in the low group were used to estimate abilities in the high group or the middle group, substantial differences in ability estimate means were found. Slinde and Linn interpreted this to mean that Rasch parameters were not really invariant and that Rasch equating procedures were not particularly useful for the problem of vertical equating.

Gustafsson (1979) criticized this interpretation. He suspected that the difference between means was due to regression artifacts which resulted from the fact that Slinde and Linn had estimated abilities and subgrouped people on the basis of only 18 of their 36 items. Individuals would not be expected to perform, in a relative sense, as extremely in either direction on the entire 36 items as they did on the easy 18; therefore, a difference between means would be expected. To support his hypothesis, Gustafsson performed a computer simulation modeled closely after the Slinde and Linn study, with the notable exception that the assumed invariance properties of the Rasch model were built in. His simulation showed that the parameter estimates obtained in the different groups were different but that this was due to a regression artifact and not to a lack of invariance. He suggested that Slinde and Linn reanalyze their data, subgrouping individuals on the basis of their total test scores.

Slinde and Linn (1979a) improved upon this idea by obtaining data from 1,638 examinees on two different tests including a 60-item reading comprehension test. The first test was used to independently subgroup examinees. The 60-item test was then split on the basis of item difficulty into two 30-item tests and their original study was essentially replicated. They found that the mean differences disappeared in comparisons of the middle with the high group (i.e., a calibration from one group applied to the other group). Whenever the low group was compared with another group, the differences persisted. This finding was attributed to the effects of guessing: No allowance is made by the one-parameter model for the possibility that correct responses can be obtained through guessing. When multiple-choice items are used, as was the case here, guessing undoubtedly occurs and probably tends to bias the results. Most likely this was a more pronounced effect for the low-ability group where examinees knew the correct answer less often and were more likely to guess.

A reanalysis of the Anchor Test Study data by the same authors (Slinde & Linn, 1979b), however, suggested a slightly different interpretation. In this study, the one-parameter model was used to vertically equate adjacent-grade pairs of published vocabulary and reading tests using an anchor-test procedure. Despite the considerable lack of model-data fit exhibited by all the tests (possibly due to multidimensionality, speededness, non-uniform item discrimination, and/or guessing), Slinde and Linn concluded that the Rasch model provided encouraging results for the problem of vertical equating. That is, differences between equated log ability estimates (computed for examinees who were administered both tests) were, typically, a fraction of the size of the standard error of measurement for either test and usually amounted to less than one raw score point throughout the ability range.

One essential difference between the two earlier studies and the later one was that the separation of high- and low-ability groups was more extreme in the earlier studies than would probably be encountered in actual grade-to-grade equating. The difference between the groups for the difficult and easy tests was five to six times greater (in terms of standard deviations on the log-ability scale) in the earlier studies than it was in the later study. Slinde and Linn (1979b) pointed out that the procedure used in the earlier studies constituted a much more severe test of the utility of the Rasch model for vertical equating. Additionally, the more recent study employed an anchor test for equating, and this procedure may significantly affect equating results.

Divgi. Divgi (1980, 1981a, 1981b) presented a series of studies investigating model fit and the applicability of the one-parameter model for vertical equating. In 1980, he devised a nonparametric goodness-of-fit test to compare IRT item calibrations. This test is relevant to IRT equating because the adequacy of the IRT equating transformation relies so heavily on the adequacy of the original item parameterizations. The test is applied by first performing two calibrations on equivalent samples of the same size, thus yielding two sets of item parameter estimates. An independent validation sample is then tested and scored twice (once with each set of parameters), resulting in two sets of ability estimates. The likelihood of the set of item responses given each theta estimate is computed. The proportion of cases (P) for which the likelihood is higher in the first calibration is used in a (binomial) test of the null hypothesis that the two calibrations provide equally good fit. This test, then, provides an indication of how well the item parameter estimates predict responses in situations where they will be applied (i.e., the validation sample).

Divgi applied his index to a study of a reading test calibrated according to the one-parameter IRT model both in a high-ability sample and in a low-ability sample (N=500 for each group). For a high-ability validation group (N=100), the high-ability calibration (with P=.86) had a better fit whereas for a low-ability validation group (N=100), the low-ability calibration (P=.06) fit better; in both cases, the probability of these results occurring by chance was less than .0001. This suggested that the two calibrations are not group-independent and, therefore, that the one-parameter model may not be appropriate for vertical equating.

Divgi (1981a) presented further data in support of his contention that the one-parameter model is not appropriate for vertical equating. First, he modified the Rasch-model fit statistic (Wright & Panchapakesan, 1969), making it more powerful. He then applied it to a national reading test, and found that while the old test rejected 16% of the items, 69% of the items were rejected by the new index. Divgi concluded that ability estimates are not item-free as the model claims. He went on to form easy and difficult subtests and to compute a theta estimate for each of more than 5,000 examinees from each subtest. The mean of the standardized difference scores was close to zero; this is typically found in Rasch-model studies and is usually presented as evidence favoring the use of the one-parameter model for vertical equating. However, regression of this difference score on an independent reading ability score (predicted by the other tests in the battery) yielded a significant quadratic relationship. The difficult subtest resulted in higher theta estimates than did the easy subtest for both low-ability and high-ability examinees. Divgi speculated that this was due to guessing on the difficult test and ceiling effects on the easy test.

Divgi (1981b) presented an alternate method for studying bias in vertically equated scales in which all examinees are tested and scored on equated tests X and Y. Bias (i.e., the difference between scores on the equated tests) is computed for each examinee. The sample is grouped on an independent measure of the ability, and mean bias is computed for each group and plotted against ability. The need for a large sample can be avoided by an approach in which bias is regressed on ability. A drawback of the method is that all persons must take both tests (a single-group design), as well as the independent measure. An advantage to using this method is that there is an absolute criterion for evaluation (bias should be zero across ability), although this is strictly true only for a perfectly reliable measure of ability. Divgi applied the method to the reading test previously calibrated by the Rasch model (N=2,000) by dividing it into difficult and easy subtests and administering it to a new sample (N=5,500). This was probably the same data reported earlier for standardized difference scores. Divgi obtained the same results: Bias was high and positive for both ability groups.

Loyd-Hoover. The adequacy of Rasch-model vertical equating was also investigated in a study by Loyd and Hoover (1980). They administered three overlapping levels of the mathematics computation test of the Iowa Tests of Basic Skills to approximately 2,000 students in grades 6, 7, and 8; each student received only one test. Level 12 was targeted to be of appropriate difficulty and content for students in grade 6, Level 13 was targeted for grade 7, and Level 14 was targeted for grade 8. Each level contained 45 test items; adjacent levels had 30 items in common and nonadjacent levels had 15 items in common. Levels 12 and 13 were administered to students in grade 6, and all three levels were administered to students in grades 7 and 8. Item difficulty parameters were estimated separately by level and by grade. The corresponding IRT ability estimates were computed and placed on a common metric. The raw scores corresponding to these ability estimates were determined; raw scores were equated by defining as equivalent those raw scores corresponding to the same ability estimate.

Three applications of vertical equating were studied: (a) adjacent test levels (12 and 13, 13 and 14) were equated when parameter estimates were obtained from two groups of comparable ability; (b) nonadjacent test levels (12 and 14) were equated when parameter estimates were obtained from two groups of comparable ability; and (c) nonadjacent test levels (12 and 14) were equated by pairwise chaining through an intermediate test level (13). Equating results were evaluated by comparing them to results obtained when two seventh-grade groups were each randomly split and Levels 12 and 13 were equated from these random samples.

The results from these applications of vertical equating were disappointing. When Levels 13 and 14 were equated twice using seventh- and eighth-grade students, respectively, Level-14 equated scores were consistently higher by one to two raw-score points for the eighth-grade students than for the seventh-grade students. Discrepancies of the same magnitude were observed when Levels 12 and 13 were equated using sixth-, seventh-, and eighth-grade students. Equated scores on Level 13 were consistently highest for eighth-grade students and consistently lowest for sixth-grade students. When Levels 12 and 14 were directly equated through 15 common items using seventh- and eighth-grade students, higher Level-14 scores were consistently obtained by the older students (the mean difference was greater than two raw-score points). Similarly, equating Levels 12 and 14 via chaining through Level 13 resulted in higher equated scores for eighth graders than for the seventh graders. Comparison of these results with the results from random splits of seventh graders indicated that these discrepancies were larger than would be expected from sample differences in item parameter estimates. Post-hoc analyses suggested that the unidimensionality assumption of item response theory had been violated.

Guskey. Guskey (1981) used the one-parameter model to vertically equate Levels 9 through 14 of the reading comprehension subtest of the Iowa Tests of Basic Skills (ITBS). This subtest contained 178 items arranged sequentially by age level from lowest to highest. Each level contained overlapping sets of items (i.e., anchors) with the levels immediately preceding and following it. Item difficulty estimates were obtained for each item separately within each level.

The mean difference between anchor-item difficulty estimates was computed for each adjacent-test-level pair. These differences were used to transform the raw scores at each test level to the metric of Level 11. The transformed ability estimates were compared to the norm-referenced grade-equivalent estimates published with the ITBS manuals. At the extreme ability levels, IRT ability estimates increased much more rapidly than the ITBS grade equivalents; the correspondence between the two sets of ability estimates was closer for the middle ability range. At the lower range of ability, however, larger differences were observed between the two scales regarding estimates on Levels 9, 10, and 11 and Levels 12, 13, and 14. That is, students taking the lower-level test forms and those taking the higher-level test forms may be assigned the same IRT ability but may differ by as much as an entire year on the ITBS grade-equivalent scale.

To investigate this score gap, Guskey collected new data from other students in this lower ability range and compared their IRT and grade-equivalent ability estimates on the reading comprehension subtest with their patterns of scores on three vocabulary and mathematics subtests of the ITBS. These supplementary analyses suggested that grade-equivalent scores underestimated ability in this range. Moreover, there were no differences between the scores on the new subtests for those two groups of students who would have been assigned the same IRT abilities but different ITBS grade-equivalents. These results could not be attributed to regression artifacts. Guskey concluded that the IRT scale was more precise and stable across the ability range.

Guskey's endorsement for the one-parameter IRT model may be justified for this data set alone. It is important to note that his samples (1,000 examinees at each of six grade levels) were randomly selected from examinees with scores between 50% and 80% correct. This strict curtailment ensures that only those examinees for whom the test level was appropriate were included. That is, no examinees were included if the test level was too easy or too difficult for them. In other words, Guskey has done no more than to test the feasibility of the one-parameter model under conditions that satisfy model assumptions. Nevertheless, practical equating situations demand that the equating transformation be applied to the entire score range, including the lower region where guessing is likely to occur. It is not unreasonable to expect the one-parameter model to perform poorly at the lower score range, since it includes no provision for guessing. An equating procedure should be selected only after its superiority across the entire score range has been demonstrated.

Holmes. Holmes (1981, 1982b) employed the one-parameter model to vertically equate sets of items selected from five reading and mathematics subtests of the Comprehensive Tests of Basic Skills. Item response data were available from approximately 6,700 third and fourth graders who took Level I of the CTBS. A principal-components analysis was performed on the tetrachoric interitem correlation matrix, and 32 items that loaded highly on only the first factor of the two-factor solution were selected. Item difficulty parameters were obtained for these items. The 20 easiest items formed one test and the 20 most difficult items formed another test. This resulted in two 20-item tests that shared eight anchor items. Grade-3 responses to the easy test and Grade-4 responses to the difficult test were used for the equating. The average difference between the anchor-item difficulty estimates from the two groups of data was computed and used to transform Grade-4 ability and difficulty estimates to the Grade-3 scale. Since all the students had actually responded to all the items in both the easy and the difficult tests, two ability estimates could be obtained for each student. The average standardized difference between these pairs of estimates was computed and used to evaluate the accuracy of the equating procedure.

Holmes found that the items fit the one-parameter model well using several different definitions of fit. Nevertheless, standardized differences between the two ability estimates computed for each student revealed that the difficult-test ability estimates were consistently higher than the easy-test ability estimates for students in the low-ability range. The results from this equating procedure were applied to 2,000 third and fourth graders in a cross-validation group. Holmes observed that students in Grade 3 received consistently higher ability estimates from the easy test whereas students in Grade 4 received consistently higher ability estimates from the difficult test.

She discussed the implications of these results in terms of out-of-level testing for selected students. The most likely candidates for out-of-level testing were high-ability third graders and low-ability fourth graders. Yet the cross-validation results implied that both these groups of students would have received lower ability estimates from the out-of-level tests than they would from the tests constructed for their specific grade level. Holmes concluded that the one-parameter model was inappropriate for vertical equating across the ability range.

Summary. Together, these studies suggest that vertical equating using the one-parameter IRT model works when (a) model assumptions are satisfied, (b) the tests are of nearly equal difficulty, and (c) the group ability levels are nearly the same (Guskey, 1981; Slinde & Linn, 1979b). Problems may result, however, if the two groups are widely different in ability or if they are of sufficiently low ability that guessing occurs with any frequency (Divgi, 1980, 1981a, 1981b; Holmes, 1981, 1982b; Loyd and Hoover, 1980; Slinde & Linn, 1979a). Unfortunately, most items used in objective tests can be answered correctly by guessing and may often be used in environments where guessing is likely to occur. The three-parameter logistic model extends the Rasch model to account for guessing and thus may be more generally useful.

Three-Parameter Model

Cook-Eignor-Petersen. Cook, Eignor, and Petersen (1982) investigated item-parameter invariance, defined as the stability across time and samples, of item parameter estimates calibrated using the three-parameter logistic model. This was primarily a linking study, but is considered here because each test was equated to itself after the items were linked. Each of several tests of various content (SAT verbal, mathematics, and achievement tests) was administered twice, at different times and to different samples of approximately 2,000 examinees each. A linear transformation was then performed to put the two sets of item parameter estimates on the same scale. Various indices were used to evaluate the adequacy of the linking procedure. The indices included: (a) scatterplots of difficulty parameter estimates and scatterplots of discrimination parameter estimates for each set of paired testings; (b) correlations between the pairs of parameter estimates; (c) means and standard deviations of each of the three parameter estimates obtained at each testing; (d) the mean of the mean absolute differences between item response functions computed using the two sets of item parameter estimates and the theta estimates from the first group tested; (e) relative-efficiency curves, using the first administration of a test as the "baseline;" and (f) true-formula-score equating of the test to itself. True formula scores were defined as

$$\xi = \Sigma\left[\left(\frac{k}{k-1}\right)P - \left(\frac{1}{k-1}\right)\right] \qquad [11]$$

with summing over items where k is the number of alternatives for the item and P is the three-parameter logistic response curve specified using the transformed parameters for the item. (The formula presented by Cook et al. reduces to this standard form.) Equating was performed by computing and pairing the true formula scores, corresponding to the same theta value, for the two administrations of the same test. This was done for a series of theta values, and the equating curve was obtained by plotting the paired true scores. This curve was compared to the ideal line for equating a test to itself, having unit slope and an intercept of zero. Residuals (the differences between the true scores) were also computed and plotted against the true score from the equated test.

All the plots of conversion lines from the true-formula-score equatings were extremely close to the ideal line. Maximum absolute residuals were less than 0.5 for all SAT-V and SAT-M tests between 25 and 60 items in length, becoming slightly larger than 1.0 for an 85-item verbal test. The achievement tests were longer (100 items) and had larger maximum absolute residual values, ranging between nearly 1.0 for Biology to nearly 2.0 for American History and Social Studies. The largest residuals were typically observed for extreme scores.

Cook et al. concluded that, although there was some instability in item parameter estimation (caused more by group ability differences and possible multidimensionality of test content than by time between testings), the effect on test scores was minimal, "not trivial, [but] well within the range of the measurement error for the test" (p. 22). Of course, when pairs of tests are equated, it would be expected that larger errors would be found than when a test is equated to itself. As Cook et al. noted, even small discrepancies in equating may accumulate over time, causing scale drift.

Holmes. Holmes (1982a) conducted a study to examine the accuracy of equated ability estimates when the equated test measures something different than the reference test. She defined the "primary trait" as that trait measured by the reference test, and the "indirect trait" as that trait measured by the second test. Data included the responses of approximately 1,000 students in each of Grades 2 and 6 to appropriate levels of the reading subtest of the California Achievement Tests (CAT), the primary measures, and the Prescriptive Reading Inventory (PRI), the indirect measures. Equating was done through anchor tests. Each item on the CAT was calibrated according to the three-parameter logistic model. Then 20 anchor items were selected from the primary measures such that they closely reflected both the content and difficulty of the tests from which they were obtained.

Four content-match categories were defined on the basis of the similarity between the objectives of the CAT and the selected PRI items. The "match" item sets included the PRI items that measured objectives identical to the CAT objectives. "Similar" item sets included PRI items that measured objectives similar to CAT. "Dissimilar" item sets measured objectives not measured in CAT. "Partial" item sets measured half of the objectives measured in CAT.

Evaluation of this classification of item sets was made by a reading specialist and found to be adequate; differences in similarity ratings among the categories, however, were not great.

Item sets containing 10, 20, or 30 items from each of the four content-match categories were selected from the PRI. This yielded 12 indirect item sets. Item parameters for the 12 indirect item sets and the 20 CAT anchor items were jointly estimated. Since IRT parameters were also available for the 20 anchor items from a prior analysis of only the CAT items, the linear relationship between the pairs of estimated difficulty parameters was used to transform the PRI item parameters to the CAT metric. The rescaled item parameters were applied to the item response data of each examinee to yield one primary and 12 indirect trait estimates for each examinee. Equating adequacy was evaluated using product-moment correlations and the root mean squared difference (divided by the pooled trait-estimate variances) between the pairs of trait estimates obtained by individual examinees.

Holmes observed that the accuracy of the indirect trait estimates increased slightly as a function of the similarity between indirect item sets and the primary measures, at least in the sixth-grade sample. Accuracy was more strongly related to the number of items in the indirect item sets. There was a bias in the equating procedure, however. The average indirect trait estimates were consistently lower than the average primary trait estimates across item-set categories for grade 2; the opposite was true for grade 6. She hypothesized that this bias arose because students with zero or perfect PRI scores were deleted from the data set before equating. Results from a randomly selected cross-validation sample were very similar.

Lord-Wingersky. Lord and Wingersky (1983) used the three-parameter logistic IRT model to compare observed-score and true-score equating of an SAT verbal test to itself through a chain of five other SAT test forms. Anchor-test equating was used throughout; at each step, two test forms and the associated anchor test were calibrated simultaneously in order to place all item parameters on a common metric. Scores below the chance level were equated according to the procedure described in Lord (1980, pp. 210-211). In this procedure, all scores below the chance level are equated using conventional linear procedures, with mean defined as the sum of the c parameters and variance defined as the sum (over items) of c times $(1-c)$; these are the observed-score statistics that would be obtained for a hypothetical group of examinees with abilities at negative infinity. Lord and Wingersky observed few differences between the two methods.

Summary. The two studies of the three-parameter model considered peripheral test-equating issues: the equating of a test to itself and the effect of varying the similarity of traits measured by equated tests. Cook et al. (1982) observed slight instability of parameter estimates which may contribute to scale drift over time; larger errors would be expected when two different tests were equated (rather than a single test equated to itself). Holmes (1982a) observed, not surprisingly, a positive relationship between the accuracy of trait estimation and the similarity between the two tests being equated. No study of vertical equating, so exhaustively examined using the one-parameter model, was reported.

Comparisons Among IRT Methods

One- vs. three-parameter models. Divgi (1980) applied his nonparametric test of fit described earlier to a comparison of the one-parameter and three-parameter IRT models. The full sample of 2,000 examinees was used to calibrate items according to each of the models. As above, high-ability and low-ability validation samples were used (N=100 each). He observed significantly better fit for the three-parameter model whether the validation group was high-ability (P=.78) or low-ability (P=.82); $p < .0001$ in each case.

One- vs. two- vs. three-parameter models. Douglass (1980, 1981) conducted a large-scale study to compare the one-, two-, and three-parameter logistic IRT models for item calibration and test equating in a college classroom situation. Data were available from the midterm and final examinations in a communications course from fall 1978 and winter, spring, and fall, 1979; N= 947, 820, 594, and 1082, respectively. Three sets of examinees were selected from the fall 1979 data on the basis of their midterm examination scores. The first set was a random split of the examinees for whom both midterm and final examination scores were available. The second set corresponded to very-high- and very-low-ability examinees who scored above and below the midterm median, respectively. The third set of low- and high-ability examinees was selected such that the mean difference between the ability groups was approximately half as large as in the second set. Separate item parameter estimates were obtained for each sample for each IRT model. Item and person parameter estimates were transformed to the scale determined by the spring 1979 final examination by means of common anchor items. IRT equating was performed on the final examination scores.

Douglass observed that the c parameters of the three-parameter model were very poorly estimated by LOGIST (Wood, Wingersky, & Lord, 1976) for these data (i.e., nearly all the c parameters were set to default values because valid estimates could not be made); he eliminated the model from further consideration. He found the one-parameter-model equatings to be very stable (i.e., to give similar results) across sample sizes of 200, 600, and 800 and the two-parameter-model equatings to be less so.

Lack of an adequate criterion of equating adequacy prompted him to equate the fall 1979 final examination to itself using the three sets of examinee subgroups discussed above. This permitted a comparison of the observed equatings with the "true" known equating line that has unit slope. While neither of the methods was uniformly best, Douglass concluded that the one-parameter model provided the more acceptable method of equating.

The stability of the Rasch equating constants based on class sections was also investigated using anchor tests containing between 7 and 37 items (on a 43-item test). Douglass computed the bias in these constants to be equal to 0.25 standard deviations of ability in the most extreme case, even with the 37-item anchor test. The Rasch calibrations were consistent from sample to sample, therefore, but incorrect.

Summary. Results from these two studies are equivocal. Divgi (1980), for example, found the three-parameter model to be superior to the one-parameter model using his nonparametric fit test. Douglass (1980, 1981), however, eliminated the three-parameter model altogether because of poorly estimated c parameters and concluded that the one-parameter model was biased but consistent.

Comparisons Between Conventional and IRT Methods

Conventional vs. One-Parameter IRT

Rentz-Bashaw. Rentz and Bashaw (1975, 1977) reanalyzed the data from the Anchor Test Study and constructed a Rasch-model-based National Reference Scale for Reading. Raw scores on 14 forms of seven standardized tests of reading vocabulary and comprehension (Grades 4, 5, 6) were then placed on this scale. They first analyzed model-data fit in several different ways and concluded there was adequate fit for equating purposes.

Pairs of tests had been administered to large samples of fourth-, fifth-, and sixth-grade students. Each pair of tests was treated as one long test for the purpose of test equating; item difficulty parameters were estimated for each item separately within a test pair. The difference in average log easiness for the two tests was used as the equating constant to adjust log ability estimates and to put the two tests on the same scale. A matrix of equating constants made it possible to place all tests/abilities on the metric defined by the vocabulary test of the Sequential Tests of Educational Progress. The log ability estimates were transformed back to equated raw scores by the following procedure. For a given raw score on the base or reference test, the corresponding ability estimate on the Rasch scale was obtained. Then the raw score on the new test corresponding to that Rasch ability was computed. In practice, there were errors involved in having to assign a raw score on the equated test that is most nearly equivalent to a raw score on the reference test. Rentz and Bashaw called them assignment errors and observed that they were larger than the errors in the equating constant and were nearly 10% of the size of the standard error of measurement in their study.

The Rasch-model equating results were compared to those obtained from equipercentile equating in the Anchor Test Study. The two sets of equated scores usually differed by only one or two raw score points; rarely was this difference as high as four points. The discrepancies observed were much smaller than the standard error of measurement for the equated tests. No absolute criterion of equating adequacy was used in this study.

Beard-Pettie. Beard and Pettie (1979) compared linear and Rasch-model equating methods using an anchor-test design for equating two forms each of two levels of two different tests. For two consecutive years, different forms of communications and mathematics basic skills tests were administered to students in Grades 3 and 5. The 1976 forms contained items that were also present in the 1977 forms, and these common items formed the anchor tests for equating. Sample sizes were larger than 5,000 for each grade and content area.

Each level of each test form was separately calibrated according to the one-parameter IRT model. Beard and Pettie checked model-data fit and the stability of the anchor-test item parameters over time, and concluded that all the tests in their study showed adequate fit to the one-parameter model. Angoff's (1971) Design IV was used to linearly equate the 1976 test forms to the 1977 forms. For the one-parameter IRT model, raw 1976 scores were converted to the Rasch ability scale; the 1977 ability level that was closest to each 1976 scale value was located and converted back to a raw score on the 1977 scale. These equated raw scores were then converted to 1977 T scores.

The results were similar across test level and content. There were only small differences between the equated scores obtained by the two procedures; the largest discrepancies occurred at the lower end of the ability scale where there were few data. For all the test pairs, the scores equated by the IRT procedure were slightly, though consistently, lower than the scores equated linearly.

Golub-Smith. Golub-Smith (1980) compared the linear and Rasch-model methods of equating scores on tests of minimum basic skills administered each year to high-school students in New Jersey public schools. Twenty-five anchor items were embedded in the reading and mathematics tests that were administered to students in each of four different grades. Golub-Smith first checked the adequacy of the fit of the data to the model and concluded that there was moderate to good fit; this was done separately for each test. The author examined the equating results of the raw scores at and around the state-mandated cut-off score. The equivalent raw scores derived from the two equating methods were very similar. Eliminating the common items whose parameters were unstable from one testing to the next resulted in different raw scores being labeled "equivalent." However, there was no evidence that this editing process provided a consistently closer or worse match with the scores defined by the linear method.

The lack of an absolute criterion of equating adequacy makes the interpretation of these results difficult. That is, eliminating the unstable item parameters changed the IRT equating transformation; one can only assume that the result was an improvement in accuracy. Comparison of the two IRT transformations, however, yielded no evidence supporting or refuting that assumption.

Conventional vs. Three-Parameter IRT

Lord. Lord (1977, 1980) demonstrated IRT-based true-score equating on two calculus tests that shared 17 anchor items and were administered to two distinct groups of examinees that differed in ability. The tests were from the College Board Advanced Placement Program and the College Level Examination Program; one test was administered to each group. Item parameters and examinee abilities were simultaneously estimated on the combined data sets according to the three-parameter logistic IRT model. The administration of common anchor items ensured a common metric for all items and abilities.

IRT-based true scores were computed for each test (by summing ICCs across items in the test), and the resulting line of relationship was compared visually to the equating lines obtained by another IRT-based method (equipercentile equating applied to the estimated observed-score distribution of the combined group) and conventional equipercentile equating of observed scores. There was close agreement between the IRT-based equating methods; the results from the conventional equating were slightly different from the IRT equatings. Lack of an absolute evaluation criterion precluded more definitive conclusions regarding the relative merits of the various equating procedures.

Lord (1977, 1980) also evaluated IRT-based true-score equating by equating an 85-item verbal section of the SAT to itself by means of an external 39-item anchor test. When a test is equated to itself, the true line of relationship between test scores is known (i.e., the scores are related by a line with unit slope and an intercept of zero). This is exactly what Lord observed when he equated the SAT test to itself after its administration to two large groups (approximately 2,800 examinees each) that differed in mean ability.

Lord (1981b, 1982b) also compared the standard errors from IRT equating with those from both linear (Lord, 1975) and equipercentile (Lord, 1981a, 1982c) equating, using two forms of the SAT-V and an external 40-item anchor test. Item parameters were estimated separately for the two groups. He observed that the IRT standard errors increased in the tails, especially at the low end of the distribution. Standard errors for the linear equating were smaller than those for the IRT equating. Standard errors for the equipercentile equating were the largest of all three methods at every score level except the lowest (where IRT was the largest). All of the standard errors were less than half the size of the standard error of measurement for the tests; most were considerably smaller.

Marco. Marco (1977) conducted a study of equating methods in which he compared three-parameter logistic IRT equating (simultaneously estimating all item parameters and setting true scores on the two tests equal) with (a) pre-equating (placing all item parameter estimates on the same metric prior to a test administration by using response data from previous administrations of the items), (b) equipercentile equating, (c) linear observed-score equating (setting observed-score means and standard deviations equal), and (d) linear true-score equating (setting true-score means and standard deviations equal). The data were two SAT-V forms (containing 40 and 85 items, respectively), both given to 5,565 examinees.

IRT equating was the standard against which the other methods were compared. The evaluative criteria included a mean squared error (MSE) index of discrepancies from the IRT equating, and also the maximum absolute discrepancy from the IRT equating across the total score range and in the mid-range only. By the MSE criterion, linear true-score equating was best (i.e., closest to IRT); the other methods were similar to each other. By the maximum-discrepancy criterion in the total range, linear true-score equating was again distinguished from the other three methods. When only the mid-range (most important to college admission decisions) was considered, both the linear true-score and pre-equating methods surpassed the remaining methods and performed equally well. Assuming that the criterion was valid, linear true-score equating was shown to be the best substitute for IRT equating, with pre-equating equally good under certain conditions. Defining the IRT-based equating transformation as the criterion for evaluating equating accuracy, however, begs the question of how well IRT methods compare to conventional methods in practical equating situations.

Bejar-Wingersky. Bejar and Wingersky (1981, 1982) investigated the adequacy of the three-parameter logistic IRT model for pre-equating the Test of Standard Written English (TSWE). Specifically, they studied the fit of the response model to two forms of the TSWE and its effect on section pre-equating. They observed some lack of model fit at the item level (by comparing observed and theoretical item-on-ability regressions) and at the subscore level where multidimensionality was evident.

Bejar and Wingersky evaluated pre-equating and IRT-based true-score equating by visually comparing the resultant equating tables to those obtained from three methods of conventional equatings used here as criterion equatings: (a) linear equating using SAT-V and SAT-M as anchors, (b) linear equating using only the SAT-V as anchor, and (c) equipercentile equating using SAT-V as anchor. The first criterion and IRT-based equatings were more discrepant for the test form which showed marked multidimensionality; the amount of discrepancy between the conventional and IRT-based equating methods was a function of the old form chosen for the equating. The IRT-based conversions resulted in higher mean scaled scores with smaller standard deviations than did the conventional equatings. Given the lack of an adequate equating criterion, Bejar and Wingersky offered cautious optimism regarding the feasibility of pre-equating as an operational equating procedure.

Modu. Modu (1982) compared three-parameter logistic IRT equating with linear and equipercentile equating when the unidimensionality assumption of IRT probably did not hold. Two forms each of 11 College Board Advanced Placement Achievement tests were administered to between 1,000 and 6,500 examinees; each form contained between 35 and 120 items. The 11 tests were then separately equated. Estimated item parameters were linked through an internal anchor test (containing 14-30 items), and true scores were estimated and equated for pairs of tests. Conventional equating methods (equipercentile and linear) were also applied in conjunction with the anchor-test design. The conventional and IRT equatings were based on separate examinee samples. Tables of equivalent raw scores obtained by the three methods for pairs of achievement-test forms showed close agreement, with discrepancies of less than one point except at the extremes where data were scarce.

Petersen-Cook-Stocking. Petersen, Cook, and Stocking (1983) investigated scale drift by comparing equipercentile, three linear, and three IRT equating procedures; each procedure was applied to SAT verbal and mathematical test data and was used to equate a test to itself through a chain of five other tests. An anchor-test design was used throughout; the three-parameter logistic IRT model and true formula-score equating were used for all three IRT methods.

In the first IRT method (concurrent calibration), the first test pair and the associated anchor test were calibrated simultaneously; the resulting item parameters were then automatically placed on a common metric. The first test was transformed to the College Board scale using previously available transformation parameters. The second test, equated to the first, was then transformed to the College Board scale as well. This calibration-and-transformation process was repeated until scores on all test forms were placed on the College Board scale.

In the fixed bs method, a single test was always calibrated along with the associated anchor test; the b parameters for the anchor-test items were then held fixed in the subsequent calibration of the second test with that anchor. This process continued sequentially, with the anchor-test b parameters from the previous calibration held fixed at any stage.

In the characteristic curve transformation method, a single test was calibrated along with the anchor test. This time, a linear transformation was applied to the a and b parameters of the second test to place them on the same scale as the first test. This linear transformation was obtained from minimizing the difference between the anchor-test true scores obtained by using the item parameters from the two calibrations of a single anchor test. This process continued sequentially until all item parameters within a chain were placed on a common metric. True formula scores, then, were automatically placed on a common metric.

The three linear methods used here included the Tucker Equally Reliable, Levine Equally Reliable, and Levine Unequally Reliable models (see Angoff, 1971, for details). For all three models, scores corresponding to the same standard score were considered to be equated to each other. The models differ in their definition of the estimated means and standard deviations; in all cases, the anchor-test scores were used to estimate the scores on the two tests for the combined group of examinees. Equipercentile equating was performed by first equating each test to the associated anchor test; test scores corresponding to the same anchor-test score were considered to be equated to each other. No smoothing was performed in either the percentile tables or the equating transformation.

Petersen et al. computed a weighted (by observed score frequencies) mean squared difference between the original (scaled) score and the equated score obtained from a specific equating method; this was done separately for each method and separately for the verbal and mathematical tests. In all cases, the equating method overestimated the criterion (original) mean. For the verbal data, the three IRT methods resulted in substantially smaller total error than did any of the conventional methods; the fixed bs method was best overall. The equipercentile method was worst overall, and the Levine Equally Reliable model was the best of the linear methods. For the mathematical data, total error was smallest for the Levine Equally Reliable model and largest for the Tucker model, and the concurrent calibration method yielded errors almost as small as the best method;

equipercentile and the other two IRT methods yielded much larger errors.

The authors noted that the content differences across test forms was greater for the verbal tests than for the mathematical tests. Also, the base verbal form was longer than the other verbal test forms; none of the mathematical tasks differed in length. Hence, it appears that linear equating methods perform adequately for reasonably parallel tests, but that IRT methods (especially concurrent calibration) performed better for nonparallel tests.

Hicks. In another study of scale drift conducted at Educational Testing Service, Hicks (1983) compared conventional and IRT methods for equating the Test of English as a Foreign Language (TOEFL) after chaining. Three conventional and three IRT equating methods were examined in this study; two sections of TOEFL were each (separately) equated.

The IRT methods included the following: (a) fixed *b*s procedure (described above), where all *b* parameters were held fixed at pretested values (*a* was limited by 0.0 and 1.5); (b) modified three-parameter, where *a* and *c* were held fixed at predetermined ("representative") values, and *b*s were re-estimated using the characteristic curve transformation described above; and (c) three parameters re-estimated, where all three parameters were re-estimated and scaled using the characteristic curve transformation (no limits on *a*). Conventional equating methods included (a) equipercentile, (b) Tucker linear, and (c) Levine linear (the authors gave no further description of which Levine method was used). All of the conventional methods estimated test-score distributions from the combined examinee group.

A separate base form was established for each of the six equating methods. Instead of equating the base-form TOEFL to itself, the last (eighth) form in the link was equated (a) to the previous form in the link and consequently back to the base form and (b) directly to the base form through common items. The "direct" equatings served as a criterion against which the "chain" equatings were compared. As in the study described above, a weighted mean difference score was computed for each method. Comparisons involving equipercentile equatings were made only over the range of observed scores.

Fixed *b*s scaling provided the least equating error for both sections of the TOEFL, followed by the modified three-parameter and the Tucker models, respectively. The Tucker and Levine linear models yielded similar results.

Conventional vs. One- vs. Three-Parameter IRT

Marco-Petersen-Stewart. Marco, Petersen, and Stewart (1980) used SAT-V data to compare the best of 40 linear methods (that varied in terms of underlying assumptions) with two equipercentile and two IRT equating methods; all methods used an anchor-test design. Conventional equating was performed two different ways: (a) directly, where scores on each test form are first equated to the anchor test; scores that are equated to the same anchor-test scores are said to be equated to each other; and (b) using frequency estimation (Angoff's Design IV), where score distributions for the two test forms for the combined group of examinees are estimated from the anchor-test-score distribution for the combined examinee group. The one- and three-parameter IRT models were both used in this study to estimate true formula scores prior to equating.

Marco et al. reported the results from two basic study designs: (a) equating a test to itself, varying the difficulty and type of the anchor test (i.e., external vs. internal) and the similarity of ability levels in the two samples; and (b) equating tests of differing difficulty using an internal anchor test, varying the sample ability levels and the spread of test difficulties. Tests equated to each other had similar content (including an equal distribution of item types) and were of equal length. For the first design, the criterion score (or the score to be estimated), was defined as the test score on the first form to which the second, identical form was equated. For the second design, the results from an "ideal criterion equating" (using data from all cases in a single-group equating) provided the criterion, or "correct," score. Two ideal criterion equatings were used: (a) an equipercentile equating of observed scores, which was biased toward equipercentile methods; and (b) an equipercentile equating of true scores estimated from the three-parameter model, which was biased toward IRT methods. The evaluative indices were based on the difference between the criterion score and the corresponding estimated criterion score for a raw-score value. Total error was the mean squared difference, weighted by the number of examinees obtaining the given raw score and standardized by dividing by the product of the criterion variance and sample size; in this manner, results could be compared across equating situations as well as models. Squared bias was the mean difference score squared and divided by the criterion-score variance. Both evaluative criteria were computed over the range of raw scores above the chance level (i.e., in the area in which IRT methods can equate) for each model.

For the first design, when the anchor test paralleled the total test, linear equating was found to be best. IRT methods rated second, regardless of differences between samples. When the anchor test was easier or more difficult than the total test, however, only the IRT models were robust for between-sample differences; linear equating performed well when the samples were similar.

For the second design, in which the equated tests were of different difficulty, the three-parameter model was best according to the IRT-based criterion; both equipercentile methods were good, the one-parameter model was poor, and the linear model was extremely error-prone. This was regardless of how similar the samples were. According to the equipercentile criteria, the equipercentile methods were best for similar samples, followed by the IRT methods, with linear equating far behind. When the samples were dissimilar, the three-parameter model was best, followed by the equipercentile methods, then the one-parameter model, and finally the linear model.

Since linear equating (or at least the best of the 40 methods tried) seemed best for equating a test to itself, linear equating probably also would work well for equating parallel tests. The best linear method was not explicitly identified, and was undoubtedly different for different parts of the study; hence, there may have been a large degree of capitalization on chance. Had only one or two linear methods been included, linear equating might not have been a clear favorite for even this limited situation. The curvilinear methods gained an advantage when the tests to be equated were nonparallel, with the three-parameter IRT equating method best for the most extreme conditions.

Kolen-Whitney. Kolen and Whitney (1982) equated 12 forms of each of five subtests of the Tests of General Educational Development (GED) using linear, equipercentile, and one- and three-parameter logistic IRT methods. One of the test forms was designated an anchor; each of the other 11 forms was equated directly to the anchor form. Each examinee was administered two anchor-form subtests and the two corresponding subtests from another form; approximately 200 examinees responded to each form of each subtest. Examinees with zero or perfect scores were deleted from the sample. A 10% hold-out sample, stratified on the basis of socioeconomic status and geographical region, was used for cross-validation (i.e., consistency) purposes.

Kolen and Whitney used Angoff's (1971) Design I (for equally reliable tests) for conventional equatings. For IRT equatings, the following procedure was used. First, all anchor-form item parameters and abilities were estimated using LOGIST (Wood, Wingersky, & Lord, 1976). The examinee ability estimates were then held fixed for the other test forms while the item parameters were estimated using LOGIST; this was done separately for each of the 11 test forms. Estimated true-score equating was used to equate scores on each test form to the anchor form; both the one- and the three-parameter logistic IRT models were used throughout. The $\underline{c}$ parameters for Forms 1 to 11 (for the three-parameter model) were fixed at the modal value of the corresponding anchor form. Scores of zero on any pair of forms were equated to each other; scores below the pseudo-chance level were equated via linear interpolation. Kolen and Whitney computed the mean squared difference (adjusted for test length) between anchor-form (integer) scores and the transformed scores on the other form with identical percentile ranks in the cross-validation distributions.

In general, linear and one-parameter IRT equating yielded the most stable results; equipercentile equating and the three-parameter model were typically much worse. The authors attributed much of these results to small sample sizes and some difficulties encountered in estimating th parameters of the more complex IRT model. Slight differences in the dimensionalities of the five subtests were not reflected in differences in equating results across subtests.

Conventional vs. One- vs. Two- vs. Three-Parameter IRT

Kolen. Kolen (1981) used the equivalent-groups design to compare linear, equipercentile, and several IRT methods for equating nonparallel tests. Subtests of an old form of an achievement test, the Iowa Tests of Educational Development, were equated to the same-named subtests of two levels of a new form: an easier level and a level of the same difficulty as the old form. Between 1,500 and 1,900 students took each test, one third being held for cross-validation and the remainder being used for equating. The IRT models used included one- (traditional and modified to permit different tests to have different *a* values), two-, and three-parameter logistic models with both estimated-true-score equating and estimated-observed-score equating (using equipercentile equating on estimated observed-score distributions). The criterion was the stability of cross-validation as indexed by the mean squared difference between raw scores on the old form and equated scores on the new form having identical percentile rank for the cross-validation sample; the smaller the index, the more stable the results.

For equating the new form's easier level to the old form, the estimated-observed-score equating for the three-parameter logistic model was definitely best. The linear equating was by far the worst. For equating the new form of a more difficult level to the old form of the same difficulty, the estimated-true-score equating for the three-parameter model was the most stable.

Although the three-parameter model was best in both situations, different procedures using the model were best for the two different situations. Kolen speculated that it may have been because he used linearly extrapolated equated scores below the chance level of *c* (others have ignored this part of the scale when computing equating criterion indices) or it may have been related to LOGIST's weakness in estimating *c*'s. He also noted that the criterion was not a measure of accuracy.

Phillips. Phillips (1983) used several different methods to vertically equate different levels of an achievement-test battery. Two tests (Reading and Math) and two grade levels (4 and 8) were studied. A "scaling" test was used throughout to place all scores on the same metric; this scaling test was essentially a single external anchor test that contained items from the full range of test-form difficulties. Equipercentile equating was used as the criterion against which the other (IRT) methods were compared; cumulative frequency curves were smoothed (unidentified method) before equating. True-score equating was used for all IRT methods; comparisons were limited to those true scores above the chance level (i.e., greater than the sum of the cs). Items were calibrated separately for each test for all IRT equating methods.

The IRT models included the following: (a) one-parameter logistic; (b) "double-modified" one-parameter, where as were permitted to vary across tests but were constant within a test, and a constant lower-asymptotic parameter was used for each item; and (c) modified two-parameter, with a constant lower-asymptote parameter assigned to each item. The more traditional three-parameter model was omitted from consideration because of the estimation problems inherent with small (300 to 500) sample sizes. Nevertheless, Phillips' modified two-parameter model is classified here as an IRT model with three parameters; similarly, the double-modified one-parameter model is considered as a two-parameter IRT model.

Mean absolute differences between equated (scaled) scores were computed for all possible pairs of methods. As a basis for evaluation, each method was applied separately to two random samples (N=500) of students at each grade level to equate a test to itself; the difference between equating transformations from a single method was used as a baseline measure of equating error.

In general, differences between methods were of approximately the same magnitude for all grades and methods. The single exception was the relatively large discrepancy between the one-parameter and equipercentile equatings for the Grade 4 Reading test. The modified one- and two-parameter models were fairly consistent with the equipercentile method throughout (two-parameter slightly more so) and were more consistent with each other than they were with the Rasch model.

Summary

Although a number of comparisons among conventional and IRT methods were made, the methodology used (i.e., test content, types of examinees, data collection design, implementation of equating methods, and, especially, evaluative criteria) were so diverse that no simple conclusion is possible regarding a best method of equating under all circumstances. Some studies (Beard & Pettie, 1979; Golub-Smith, 1980; Modu, 1982; Rentz & Bashaw, 1975, 1977) found no differences among methods, while others (Bejar & Wingersky, 1981, 1982; Hicks, 1983; Kolen, 1981; Kolen & Whitney, 1982; Lord, 1977, 1980; Marco, 1977; Marco et al., 1980; Petersen et al., 1983; Phillips, 1983) found that the methods ordered themselves differently depending upon the conditions under which equating was performed and the results were evaluated. Some conclusions can be drawn by considering these results in light of the dimensions of equating needs.

Relevance of Previous Research to Practical Equating Situations

In the studies cited above, individual power tests were equated to each other; no study attempted to equate speeded tests. Most of the tests were assumed to be unidimensional, and checks for multidimensionality were performed only occasionally. The discussion of the literature as it applies to the practical equating needs is perhaps best done within the parallel/non-parallel test distinction.

Equating Parallel Tests

Theoretically Appropriate Methods

Conventional and strong-true-score methods of equating are appropriate whenever individual tests to be equated are parallel; IRT methods are appropriate under the added constraint that the tests are unidimensional and not speeded. Empirical comparisons of these equating methods yielded results that were consistent with expectations. That is, when the assumptions underlying the equating procedures were met, few differences among the various procedures were observed.

Previous Research

Garcia-Quintana and Johnson (1979) found few differences among the conventional linear methods they investigated; Lord and Wingersky (1983) drew the same conclusion regarding true-score and observed-score IRT equating. Using an anchor-test design, Marco et al. (1980) observed that linear methods worked better than IRT or equipercentile methods when a test was equated to itself and the two samples were similar in ability. Similarly, Petersen et al. (1983) found that linear methods worked better than equipercentile or IRT methods for equating a test to itself through a chain of other tests. Presumably, linear methods would also work best for equating parallel tests as long as the abilities of the two groups were similarly distributed. Lord (1981a, 1981b, 1982b, 1982c), for example, found that the standard error of equating was smaller for linear equating than it was for equipercentile equating; the standard error of IRT equating was between these two values.

When test data fit the one-parameter model, few if any differences were observed between linear and Rasch equating procedures (Beard & Pettie, 1979; Golub-Smith, 1980; Kolen & Whitney, 1982). In a study comparing the three IRT models, Douglass (1980, 1981) found the one-parameter model to yield results that were more stable but also more biased than those of the two-parameter model; problems estimating the c parameter caused him to ignore the three-parameter model altogether. Kolen and Whitney (1982) found the linear and Rasch methods to be more stable than equipercentile or three-parameter IRT methods.

One of the IRT methods studied by Hicks (1983) outperformed conventional methods in terms of scale stability; the linear methods were very similar. The three-parameter IRT model performed well when a test was equated to itself (Cook et al., 1982; Lord, 1977) and, by inference, to strictly parallel tests. Modu (1982) observed few differences among linear, equipercentile, and three-parameter IRT equating methods for tests that were probably multidimensional.

In one of the few studies that reported explicit evidence of model-data misfit, Bejar and Wingersky (1981, 1982) noted that the discrepancy between conventional equating methods and IRT-based true-score equating and pre-equating was greatest for the test form which exhibited marked multidimensionality. The IRT methods produced very similar results, as did the three conventional methods.

Conclusions

Previously reported data seem to indicate that conventional and IRT procedures yield essentially the same results when they are used to equate parallel tests. The IRT procedures, however, may be more appropriate when samples differ greatly in ability (Marco et al. 1980); their superiority has not been established for multidimensional tests (cf. Bejar & Wingersky, 1981, 1982).

Equating Nonparallel Tests of Equal Difficulty

Theoretically Appropriate Methods

Theoretically, only STST methods are appropriate for equating nonparallel tests in every situation; nonparallel tests may also be equated using IRT techniques as long as the tests are not multidimensional in nature or administered with a strict time limit (i.e., speeded). Nevertheless, investigators have examined the applicability of conventional as well as IRT techniques to situations involving nonparallel tests. None have compared these procedures to STST.

Previous Research

Bianchini and Loret (1974) concluded that equipercentile methods yielded more consistent results for equating nonparallel tests than did linear equating methods, with consistency defined as the similarity of results between half- and whole-sample equatings. Linear and equipercentile methods of equating yielded virtually identical results in the study reported by Stock et al. (1980).

Lord (1977, 1980) observed that equipercentile equating of observed formula scores yielded results somewhat different from those based on IRT equating procedures; the two IRT methods, however, yielded very similar results. Kolen (1981) observed that equating methods based on the three-parameter model were more stable (in terms of cross-validation) than were other IRT and conventional methods. Similarly, three-parameter IRT methods worked better than conventional methods for equating nonparallel tests in the study by Petersen et al. (1983).

Holmes (1982a) systematically varied the degree of nonparallelism across equatings and studied its effects on the accuracy of IRT equating. Although the experimental manipulation was not strong and differences across types of tests were not great, her results suggested that equating accuracy may be affected by the similarity of item content in the tests to be equated.

Conclusions

No definitive conclusions can be drawn from the literature regarding which equating method is best applied to nonparallel tests of equal difficulty. This is in large part due to the lack of an adequate criterion for evaluating the results. Some researchers compared results across methods and merely looked for differences in the equating transformations (Lord, 1977, 1980; Stock et al., 1980); at best, equating methods were compared for consistency (Bianchini & Loret, 1974) or cross-validation stability (Kolen, 1981). When differences among methods were observed, it was not clear which, if any, of the methods was more accurate. Data concerning the stability of equating results are only marginally relevant; Douglass (1980, 1981), for example, found the Rasch model to be very stable but also inaccurate for equating parallel tasks.

Whereas some researchers observed differences between the conventional and IRT equating methods (Bianchini & Loret, 1974; Lord, 1977, 1980; Kolen, 1981; Petersen et al., 1983), others (Stock et al., 1980) did not. In view of Holmes' (1982a) results, it is possible that observed differences across methods may be a function of the degree to which the tests being equated were nonparallel or multidimensional. It is not known, for example, to what extent the calculus tests used by

Lord (1977, 1980) were parallel or unidimensional. They were, no doubt, closer to being parallel than were the GRE subtests and the Miller Analogies Test reported by Stock et al. (1980). Little can be concluded without further research.

Equating Tests of Different Difficulty

Theoretically Appropriate Methods

As with nonparallel tests of equal difficulty, tests of unequal difficulty are appropriately equated only using STST methods; unidimensional power tests of unequal difficulty may also be equated using IRT. Researchers have typically employed equipercentile and IRT methods in their investigations of vertical equating.

Previous Research

It appears that the vertical equating of tests is a much more difficult task than is the equating of tests that are similar in difficulty. Slinde and Linn (1977), for example, rejected equipercentile equating as a viable method for vertical test equating; the one-parameter IRT model was similarly rejected by the same authors in later studies (1978, 1979a; cf. Gustafsson, 1979). However, Slinde and Linn (1979b) later changed their opinion and suggested that the one-parameter model may be suitable for vertical equating with an anchor test when the groups are not widely different in ability.

Rentz and Bashaw (1975, 1977) and Guskey (1981) reported successful applications of the one-parameter model for the problem of vertical equating. Holmes (1981, 1982b) and Loyd and Hoover (1980), however, found serious evidence of bias in their data sets and cautioned against the use of that IRT model for vertical equating. Divgi (1981a, 1981b) reached the same conclusion. Similarly, Phillips (1983) found the traditional one-parameter model to yield equating results discrepant from other IRT and conventional methods; the one- and two-parameter models, when modified to permit non-zero lower asymptotes, yielded results consistent with equipercentile equating methods.

Conclusions

The results concerning the vertical equating of nonparallel tests appear to be equivocal. Conventional methods do not appear to be adequate. Although some researchers suggest that the (unmodified) one-parameter IRT model may be appropriate for vertical equating under certain circumstances, the possibility of scale bias precludes enthusiastic endorsement of that method. It appears that some provision for a pseudo-guessing parameter needs to be included in an IRT model before it is appropriate for vertical test equating.

The Criterion Problem

Previous Approaches

The relative merits of the various equating procedures may be obscured by the lack of a criterion for evaluating equating accuracy. In general, researchers have evaluated their equating procedures in one of the following ways: (a) looking at discrepancies across methods, (b) computing indices of consistency and/or stability, (c) equating a test to itself, and (d) comparing equated scores to observed scores when all examinees respond to all test forms.

Discrepancies Across Methods

Many of the studies (Beard & Pettie, 1979; Bejar & Wingersky, 1981, 1982; Garcia-Quintana & Johnson, 1979; Golub-Smith, 1980; Guskey,1981; Lord, 1977, 1980; Loyd & Hoover, 1980; Modu, 1982; Slinde & Linn, 1977; Stock et al., 1980) compared the results of different equating methods simply by examining tables of equivalent scores (and sometimes their plots) to see whether the different equating methods resulted in the same equated scores. Typically, this led to a qualitative statement such as noting that there were small discrepancies (e.g., one raw score point or less) throughout most of the raw-score range but larger discrepancies in the chance-score range. These discrepancies were sometimes compared with the standard error of measurement of the equated tests (Rentz & Bashaw, 1975; Slinde & Linn, 1979b) to evaluate the seriousness of these differences.

In every case, even if there were no discrepancies between equating tables from two methods, the most that could be said was that neither method was better than the other. In some cases, this was the desired conclusion. Jaeger (1981), for example, developed several indices to identify the conditions under which linear equating could be substituted for equipercentile equating with no change in results. In most cases, however, it would be desirable to be able to specify which of two equating methods is better when the results are discrepant and to determine the amount of error in equating. This is the familiar criterion problem, and no completely satisfactory index has been proposed. Lord's (1975, 1981a, 1981b, 1982b, 1982c) standard-error-of-equating indices allow comparisons of equating quality at different score levels, but do not solve the practical problem of knowing the correct or best equated score. This was illustrated in the study described above (Lord, 1981a, 1982c) in which the equated scores from two methods were more than a standard error of measurement apart.

Alternatively, when various equating methods are being compared, one method can be designated a "criterion" or standard equating against which to compare the other equatings. Hicks (1983) investigated scale drift after a chain of equatings by comparing the final equating transformation with that obtained in an anchor-test equating. In a similar approach, Marco (1977) computed the mean squared difference between the scores from each equating method and a criterion IRT equating across test-score values in the equating table. In a later study (Marco et al., 1980), he evaluated the equatings based on various population subsamples by comparing them to two criterion equatings (equipercentile and IRT) derived from the total sample. The indices he used (total error and bias) were both based on the difference between the criterion equated score and the equated score obtained by a method studied in a subsample.

The problem with this approach is that it assumes what it wants to show, namely which equating method is in some sense "best." Unfortunately, that is the problem facing any attempt to compare equating results obtained by the anchor-test or equivalent-groups methods. A criterion of equating accuracy is needed that involves more than a mere comparison of the similarity of equating results.

Consistency and Stability Indices

Kolen (1981) and Kolen and Whitney (1982) avoided defining an absolute criterion and examined instead the stability of equating results when applied to a cross-validation sample. In both studies, stability was defined as the mean squared difference between the raw score obtained on the old form and the equated score on the new form computed for a cross-validation sample. Similarly, Bianchini and Loret (1974) compared equating methods by computing a root-mean-squared-error index that was based on the discrepancies between equivalent scores from half- and whole-sample equatings.

These approaches provide information concerning which of the equating methods yields the equating transformation that is most stable across samples of examinees. These indices do not, however, identify the most accurate transformation in an absolute sense.

Equating a Test to Itself

An absolute index of error exists for the trivial case in which a test is equated to itself. A given score from the first administration of a test should correspond to the same (equated) score from the second administration. Therefore, a plot of the equating transformation should be a straight line with zero origin and unit slope. The discrepancy between the observed transformation and the "ideal" line is an index of equating accuracy.

Several studies reviewed here equated a test to itself in order to demonstrate goodness of equating (Cook et al., 1982; Douglass, 1980, 1981; Lord, 1977; Marco et al., 1980). This situation may be considered as a case of equating two strictly parallel tests. As such, it provides little information concerning how well an equating method works when nonparallel tests need to be equated.

A test was equated to itself non-trivially by Petersen et al. (1983) and Lord and Wingersky (1983). Those studies focused on equating errors that occur when several test forms are equated and chained together. Thus, the errors that were observed between the original test scores and the equated scores (after chaining) provided a bona fide index of scale drift.

Discrepancies Between Observed Scores and Equated Scores

When the single-group design is used (i.e., when test scores for all examinees are obtained on all tests), there exists some "absolute" criterion of equating accuracy. For each examinee, the equated score from the second test should be identical to the score he or she received on the first test. The discrepancy between observed and equated scores can be readily computed as a measure of equating accuracy. Several indices based on these discrepancies have been proposed. All these indices are based on the difference between two scores that are considered to be equated, and all yield a value of zero for errorless observed-score equating.

The most prominent index is the standardized difference (between two ability estimates computed for each examinee from two sets of item parameter estimates) frequently used in Rasch-model studies of vertical equating (Divgi, 1981a; Gustafsson, 1979; Holmes, 1981, 1982b; Slinde & Linn, 1978, 1979a). Similarly, the bias indices developed by Divgi (1981b) and Yen (1982) and the root-mean-squared-difference index of Holmes (1982a) are also applicable in studies using the single-group design. In almost every case, however, the amount of measurement error overwhelms the amount of equating error that is present in any equating transformation.

It is important to note that these indices are based on observed scores. That is, the equating transformation is applied to an examinee's observed score on the new test in order to arrive at an equated old-test score. The difference between an examinee's observed score on the old test and the examinee's equated score is then computed. The process repeats for each examinee. This process is appropriate if the purpose of the transformation is to equate observed test scores. Since the examinees' true scores can never be known, Lord (1982a) states that it is appropriate; Braun and Holland (1982) and Rubin (1982) concur.

Observed-Score vs. True-Score Equating

Observed-score equating may be inappropriate for several reasons. First and foremost, strict equating requires that the observed scores and the corresponding transformed scores have identical frequency distributions in all groups tested (Lord, 1977, 1980, 1982a). In general, this requirement will not be met if the equating transformation is based on (fallible) observed scores (Lord, 1977, 1980, 1982a). Moreover, following Lord's requirement of equity, where it makes no difference to the examinees which test they are given, it becomes clear that tests that differ in reliability and/or difficulty cannot be equated. One is faced with the following paradox: "...scores $\underline{x}$ and $\underline{y}$ on two tests cannot be equated unless either (1) both scores are perfectly reliable or (2) the two tests are strictly parallel" (Lord, 1980, p. 198). In other words, it is appropriate to equate observed test scores only when it is impossible or unnecessary to do so.

On the other hand, one can attempt to equate true scores instead of observed scores, following Morris' (1982) definition of weakly equated tests, where "each individual in the test population has the same expected score on both tests" (p. 171).

True-score equating satisfies the requirements of group invariance and equity. Moreover, true scores can, theoretically, be equated even for tests that are not strictly parallel (i.e., for all practical equating situations). IRT and STST provide ways of estimating the equating transformation between two sets of true scores.

The major criticism of true-score equating is that examinees' true scores are, of course, never known to the examiner. They can, at best, only be estimated from item response vectors. These estimates are then just another sort of fallible observed scores and, strictly speaking, cannot be equated. Even though the transformation itself can be estimated (using IRT or STST), one cannot substitute estimates for true scores and expect strict equating to hold. The problem, as described by Lord (1977, 1980), is that there is thus no truly appropriate way to make use of the true-score equating transformation. As he states, "Either the exact true-score equating can be used with observed scores, or else an inexact observed-score equating can be used. The real problem is that we have no criterion for choosing" (1977, p. 133). It may be interesting to note, however, that in the only study explicitly designed to compare observed- and true-score equating, Lord and Wingersky (1983) observed few differences between the methods when a test was equated to itself through a chain of other tests.

A More Satisfactory Approach

It is clear from the definition of weakly equated tests that the goal of equating is to provide a transformation for making true scores on two tests in some sense equivalent. The criterion for choosing between observed-score equating methods and true-score methods applied to observed scores also becomes clear: The evaluation of an equating transformation should be based on true scores. That is, the equating transformation, however obtained, should be applied to an examinee's true score on the new test; the equated true score should then be compared with that examinee's true score on the old test.

This evaluation process can be performed, of course, only in the situation where true scores are known for each examinee, that is, in a Monte Carlo simulation study. The conclusions thus derived regarding the relative merits of the various equating procedures should then be directly applicable to practical equating situations provided that the simulation procedures accurately model real-world conditions.

Design Issues for a Study of Equating

This review of the equating literature provides a basis for the design of an equating study using Monte Carlo simulation techniques. Previous studies were examined to determine those characteristics of real data that should be modeled in a simulation. Few of the studies dealt with such issues as the examinee sample size necessary for a stable equating, however. Nearly all of the studies used real data from standardized tests and available samples, with some information not explicity given and therefore not amenable to simulation. Only one of the studies formed composite scores before equating. The few simulation studies were not directly relevant to the kind of study considered here.

Examinee Ability Distributions

The assumption is often made that ability is normally distributed in the population, and therefore normally or at least unimodally distributed in a random sample from that population. The score distributions for tests measuring that ability are also assumed to be normally distributed. No data were given in these papers that contradicted the assumption of normality, although this issue was seldom directly addressed. A simulation study based on normally distributed abilities and test scores may be reasonable. It is probably much more appropriate, however, to explicitly examine distribution shapes in samples similar to those found in military situations.

Test Structures

The real tests used in the studies reviewed here were mainly standardized tests with national norms: primarily reading or mathematics tests at the elementary or high school level or admissions or achievement tests at the college or graduate school level.

Three simulation studies were reviewed. Marks and Lindsay (1972) assumed that true test scores were normally distributed, and modeled each observed score as the simple sum of a true score and a random error component. Gustafsson's (1979) simulated tests were composed of nine items at each of four evenly spaced difficulty levels. Yen (1982) designed her tests so that the difficulty parameters were approximately normally distributed; mean test difficulty was varied by adding or subtracting a constant from every item difficulty parameter in the set.

A realistic simulation should start with item parameter estimates similar to those from tests to be equated in practice. This would ensure that the results and conclusions would be as applicable as possible to a real situation. Both power and speeded tests should be simulated.

Sample Sizes

Marks and Lindsay (1972) found that 250 examinees were necessary for adequate equating, but their study used an inappropriate equating method (regression). Douglass (1980, 1981) found sample size (200, 600, and 800) to be an unimportant factor for Rasch-model equatings, but an important variable influencing the consistency of two-parameter-model equatings. Similarly, Kolen and Whitney (1982) suggested that 200 examinees may be too few for three-parameter IRT equating. Yen's (1982) sample-size variable (1,000 vs. 2,000) had no effect on equipercentile equating. The remainder of the studies involved data from national testing programs with such large numbers of examinees that sample size was no longer an issue. Typically at least 1,000 and often several thousand examinees were used for each equating. Such large numbers are probably adequate, but these studies leave unanswered the question of how many examinees are needed for equating and how this minimum sample size varies across equating methods.

Composites

Stock et al. (1980) used a GRE composite as one total test score to be equated. In all other studies, the equating methods were applied solely to individual tests. Even when a test battery was available, individual tests were separately equated. Completely unresolved are the problems of how to combine correlated subtests into a composite and how to equate composites composed of same or different tests.

METHOD

Project Overview

Equating procedures are best compared when (a) either all the relevant test-model assumptions are met or the extent of their violation is known, (b) characteristics of the testing situation are systematically manipulated, and (c) there exists a criterion indexing the accuracy of the equating transformation. In the present study, different testing situations were simulated. First, examinee responses to unidimensional parallel tests were generated; equating methods were applied to individual subtests and to composites of these subtests for different samples. Since the test-model assumptions were satisfied, the equating methods were compared and evaluated under ideal conditions.

Next, the test-model assumptions were violated in ways that modeled actual testing conditions. The parallel-test assumption of the conventional equating procedures was violated by equating subtests and composites of different lengths (i.e., different reliabilities) and different difficulties. Specifically, current Air Force equating needs and conditions were simulated. Item responses were modeled on the four AFQT subtests and examinee groups that resembled the current military applicant population. Subtest length and difficulty were systematically varied.

Two different raw score composites were computed: (a) an AFQT composite analogous to the current AFQT, formed by weighting and summing across the four subtests, and (b) a power-test composite formed from the three power AFQT subtests. These scores were equated using different data collection designs, testing models, and transformation forms. In addition, power and AFQT composites were formed by weighting and summing across already-equated subtest scores. Composite scores were also equated indirectly (as described below) using score statistics and intercorrelations from individual subtests.

The use of a Monte Carlo simulation in this study permitted a clear evaluation of the equating procedures. In a Monte Carlo simulation study, examinee ability levels and, hence, true scores on all tests can be specified a priori. Thus, the relationship between the two sets of true scores is known. The equating transformation computed from the fallible observed scores can then be compared with the known relationship between the true scores. Discrepancies between the equated and true scores on a test can be used to evaluate equating accuracy. This type of comparison is not possible when real data are used.

The equating procedures employed in the simulations were also applied to real Air Force data. That is, the same data collection designs and equating transformations that were used in the simulations were applied to item response and test score data from real examinees.

Table 1 presents an overview of the project. Equating was performed for the combinations of equating method, data collection design, and tests and composites marked. Within each cell, both parallel and nonparallel tests and composites were equated using various combinations of test length and difficulty.

Table 1
Application of Equating Methods and Data Collection Designs to Subtests and Composites

Equating method and design	Subtests*				Composites						
					Direct			Equated Subtests		Indirect	
	PC	AR	WK	NO	Power	AFQT	AFQT to power	Power	AFQT	Power	AFQT
Linear											
Single group	X	X	X	X	X	X	X	X	X	X	X
Equivalent groups	X	X	X	X	X	X	X	X	X	X	X
Anchor test	X	X	X	X	X		X	X	X		
Equipercentile											
Single group	X	X	X	X	X	X	X	X	X		
Equivalent groups	X	X	X	X	X	X	X	X	X		
Anchor test	X	X	X	X	X		X	X	X		
IRT											
Single group	X	X	X					X			
Equivalent groups	X	X	X					X			
Anchor test	X	X	X					X			
STST											
Single group	X	X	X	X	X			X	X		
Equivalent groups	X	X	X	X	X			X	X		
Anchor test	X	X	X	X				X	X		

*Subtests included: (a) Paragraph Comprehension (PC); (b) Arithmetic Reasoning (AR); (c) Word Knowledge (WK); and (d) Numerical Operations (NO).

Sample Characteristics

Three data collection designs (single group, equivalent groups, and anchor test) were investigated in this study. To implement these designs, three distinct examinee groups were generated. First, a parent distribution of examinee abilities was defined. This distribution of abilities was defined to be comparable to the distribution of abilities of current military applicants. Sample X consisted of simulated examinees whose abilities were randomly sampled from this distribution. This sample was used for all three data collection designs. Sample Y consisted of simulated examinees whose abilities were randomly sampled from the same parent distribution of ability. Sample Y was used for the equivalent-groups and anchor-test designs.

The equivalent-groups data collection design assumes that the groups administered the two tests are, in fact, random samples from the same parent population. The anchor-test design makes no such assumption and, therefore, may be more appropriate for use with nonequivalent groups of examinees. Consequently, a third examinee group (Sample Z) was generated. The mean of the parent ability distribution was increased, and examinee abilities were randomly sampled from this new distribution. Sample Z was used for both the equivalent-groups and anchor-test designs.

All test forms were equated using two different sample sizes (1,000; 2,400). In addition, a third sample size was used for Samples X and Y. Responses to a separate selection composite were generated for Samples X and Y, and the examinees with the highest scores on this composite constituted the selected sample. Selected samples ($\underline{N}$ = 1,600) were used to evaluate the equating procedures when applied to tests administered to samples of selected recruits.

Test Characteristics

To perform this evaluation, two different difficulty levels and two different test lengths were simulated for each subtest. Thus, eight different forms of each subtest were required; Table 2 presents a summary of these test-form characteristics. Even-numbered forms were the "new" tests or composites that were equated to the odd-numbered "old" forms.

Table 2
Test-Form Characteristics

Form number	Difficulty level	Length
1	easy	short
2	easy	short
3	easy	long
4	easy	long
5	difficult	short
6	difficult	short
7	difficult	long
8	difficult	long

These test forms were paired in nine distinct ways. Table 3 presents a summary of these test-form pairings. It also specifies the difficulty level of the anchor test that was used for applying the anchor-test design to each pairing.

Table 3
Test-Form Pairings

"Old" test		"New" test		Anchor-test difficulty
Form	Characteristics	Form	Characteristics	
Parallel pairings				
1	easy, short	2	easy, short	easy
3	easy, long	4	easy, long	easy
5	difficult, short	6	difficult, short	difficult
7	difficult, long	8	difficult, long	difficult
Nonparallel pairings				
1	easy, short	6	difficult, short	medium
3	easy, long	8	difficult, long	medium
3	easy, long	2	easy, short	easy
7	difficult, long	6	difficult, short	difficult
3	easy, long	6	difficult, short	medium

In the first four of these pairings, parallel tests were equated to each other. For example, an easy short test was equated to another easy short test. Nonparallel tests were equated in five different ways. For example, a difficult short test was equated to an easy short test, and a difficult long test was equated to an easy long test.

Data Collection Designs

Three separate data collection designs were evaluated in this study: the single-group, equivalent-groups, and anchor-test designs. For the single-group design, responses to both sets of test items were generated for a single group of examinees. Data from Sample X were used for the single-group design. The equivalent-groups design differed from the single-group procedure in that a different set of abilities was drawn in order to generate item responses on the new test(s). Old-test responses were always generated for Sample X examinees. New-test responses were generated for Sample Y examinees for all subtests (and for Sample Z examinees for the Paragraph Comprehension subtest). In the anchor-test design, item responses to the old test/composite were generated for Sample X. Item responses to the new test/composite were generated for Sample Y (and for Sample Z for the Paragraph Comprehension subtest). Responses to the common set of anchor items were generated for all examinee groups. Sample pairings for the anchor-test design were identical to those in the equivalent-groups design.

Individual subtest scores were equated using this project design. Because anchor-test equating can be used to directly equate composites of power and speeded tests (e.g., AFQT composites) only if separate power and speeded anchor tests are administered and later combined into an anchor-test composite, and because this procedure is cumbersome and unlikely to be used in practice, only power composites (i.e., not AFQT composites) were directly equated using the anchor-test data collection design.

Each data collection design was used with all testing models and transformations and with all combinations of test length, test difficulty, and sample size. The combinations of data collection designs and sample ability distributions are presented in Table 4. The X-X and X-Y sample pairings were used for all tests and composites; the X-Z pairing (varying ability level across samples) was used for all equatings of the Paragraph Comprehension subtest.

Table 4
Combinations of Data Collection Designs and Sample Ability Distributions

Data collection design	"Old" test			"New" test		
	Sample	Size	Ability level	Sample	Size	Ability level
Single group	X	1000	current	X	1000	current
	X	1600	selected	X	1600	selected
	X	2400	current	X	2400	current
Equivalent groups	X	1000	current	Y	1000	current
	X	1600	selected	Y	1600	selected
	X	2400	current	Y	2400	current
	X	1000	current	Z	1000	increased
	X	2400	current	Z	2400	increased
Anchor test	X	1000	current	Y	1000	current
	X	1600	selected	Y	1600	selected
	X	2400	current	Y	2400	current
	X	1000	current	Z	1000	increased
	X	2400	current	Z	2400	increased

Note. Adjectives describe the sample in relation to the parent sample (e.g., "increased" indicates that an additive constant has been applied to the mean of the parent -- "current" -- population before sampling).

Data-Generation Procedures

True subtest abilities for each examinee were sampled from a multivariate nonnormal distribution that was defined to be similar to the multivariate distribution of subtest abilities of current military applicants. This multivariate distribution was defined by the first four (marginal) moments of each subtest and by the intercorrelation matrix of the subtest scores. The true IRT item parameters for the three power subtests were sampled from a multivariate nonnormal distribution of parameters defined to be similar to the distribution of item parameters in the current ASVAB subtests. Similarly, characteristics of the speeded tests were modeled after the speeded subtest in the current AFQT.

The true abilities and item characteristics were combined with a random process to yield item responses and, subsequently, fallible observed scores for each subtest and composite. All equating transformations were derived from these observed scores and responses. Details concerning the ability distributions, parameter distributions, and specific data-generation procedures are given below.

Examinee Characteristics

Specification of the Moments of the True-Ability Distributions

Table 5 presents the summary statistics used to specify the multivariate nonnormal distribution of true abilities.

Table 5
Summary Statistics Used to Specify Multivariate Distribution of True Abilities

						Corrected correlations			
Subtest	Mean	Variance	Skewness	Kurtosis	Reliability	PC	AR	WK	NO
PC	0.090	0.795	0.170	-0.672	.80	-			
AR	0.094	0.805	0.164	-0.607	.91	.83	-		
WK	0.086	0.854	0.177	-0.860	.92	.94	.80	-	
NO	0.696	0.041	-0.455	-0.323	.91	.64	.70	.61	-

Note. Table entries adapted from Vale et al. (1981) and Ree et al. (1982). Summary statistics for the NO subtest are expressed on a proportion-correct metric. All other subtest statistics are expressed on an IRT theta metric. A constant of 0.25 was added to the mean for each power subtest before abilities were sampled for Sample Z. Similarly, 0.04 was added to the mean of the speeded (NO) subtest for Sample Z. Correlations have been corrected for the unreliability of the tests.

Power-test abilities. Table 8 in Vale et al. (1981, p. 56) presents summary statistics for modal Bayesian ability estimates derived from the responses of 500 military applicants to an experimental form of the ASVAB-8 subtests during 1978. The first four moments for the Arithmetic Reasoning (AR) and Word Knowledge (WK) subtests were used to define the first four moments of the true abilities for these subtests. Because the data from the Paragraph Comprehension (PC) subtest were not presented in Table 8, the median moments across all subtests were used as estimates for the moments of the PC abilities.[1]

Speeded-test abilities. An examinee's number-correct score on a test administered with a strict time limit is a joint function of the speed and the precision with which the items are answered. These characteristics can be respectively defined for a test with a time limit by (a) the number of items attempted (i.e., speed), and (b) the proportion of correct responses computed from the number of items attempted (i.e., precision). For a pure power test, the number of items attempted is equal to the number of items on the test; for a pure speeded test, the proportion of correct responses computed from the number of items attempted is 1.00. A time-limit test (i.e., a partially speeded power test) can be considered to be a combination of a pure power and a pure speeded test. For a partially speeded test such as Numerical Operations (NO), the values for these two characteristics lie somewhere between the limits of the pure power and pure speeded tests. The values used in this study were determined as follows.

First, the item responses to the Numerical Operations (NO) subtest from 15,115 Military Enlistment Processing Stations (MEPS) examinees who took ASVAB Forms 8, 9, and 10 were obtained from D. J. Weiss (personal communication, September 10, 1982). Because the NO subtest was administered with a time limit, not all examinees responded to every item. For each examinee, the number of items attempted was defined to be equal to the sequence number of the last item for which there was a response. All succeeding responses were coded "not reached." Missing responses prior to this point were coded "omitted." Thus, speeded-test responses were coded as correct, incorrect, omitted, or not reached. Data were analyzed only for those 14,460 examinees who omitted fewer than two items before the time limit was reached. Omitted responses were recoded as incorrect. A proportion-correct score was computed for each examinee. The summary statistics in Table 5 were based on these proportion-correct scores.

The distribution of the number of items as a function of the proportion-correct score was also obtained from these data. This distribution was used later to generate speeded-test item responses.

---

[1] The standard deviations presented in Vale et al. (1981) were actually treated as variances when the multivariate ability distribution was specified. This caused the simulated abilities to have a larger variance than should have been the case otherwise.

Varying ability. The parent ability distribution was modified before examinees were drawn for Sample Z. This modification served to simulate the difference in mean ability that might occur between military applicants in successive years, or between the group of current applicants and the mobilization population. A constant was added to the mean of the ability distribution for the three power tests and to the mean of the distribution of proportion-correct scores for the speeded test. These constants were determined as follows.

First, data were obtained concerning the distribution of applicants across AFQT categories for two successive years (R. S. Massar, personal communication, January 25, 1983). These data were collected between October and December, 1981, and between October and December, 1982. A continuous frequency distribution for each year was formed by interpolating between the midpoints of each score interval. Table 6 presents these data.

Table 6
Distribution of Applicants Across AFQT Categories

AFQT category	Score interval (percentile)	Oct-Dec 1981 Proportion Raw	Oct-Dec 1981 Proportion Cum.	Oct-Dec 1982 Proportion Raw	Oct-Dec 1982 Proportion Cum.
I	93-99	.026	.999	.034	1.000
II	65-92	.260	.973	.311	.966
IIIa	50-64	.154	.713	.172	.655
IIIb	31-49	.202	.559	.213	.483
IVa	21-30	.139	.357	.128	.270
IVb	16-20	.081	.218	.064	.142
IVc	10-15	.082	.137	.052	.078
V	01-09	.055	.055	.026	.026
N of cases		127,188		92,817	

Note. These data are for non-prior-service male applicants (first ASVAB administration) only. Data were provided by R. S. Massar (personal communication, January 25, 1983).

The 1982 applicants scored higher, on the average, than did the 1981 applicants. In fact, the 50th percentile for the 1982 applicants corresponded to approximately the 57th percentile of the 1981 applicants. According to Table 8 in the report by Ree, Mathews, Mullins, and Massey (1982), these percentiles correspond to (interpolated) AFQT raw scores of approximately 75.5 and 80.2, respectively. This raw-score difference was fairly constant throughout the ability range. Comparison of these scores with the standard deviations (for all six ASVAB forms) reported in Table 7 of that same report revealed a standard-score difference between 0.23 and 0.25. Consequently, a constant of 0.25 was added to the mean ability of the parent distribution for Sample Z. Accordingly, a constant of 0.04 (approximately 0.25 standard deviation on the proportion-correct metric) was added to the mean of the distribution of proportion-correct scores for Subtest NO.

Specification of the Correlations Among True Abilities

As just discussed, the moments of the distributions of true abilities were taken from Vale et al. (1981). That report, however, did not report correlations among the subtest ability (i.e., theta) levels, nor were any such data available elsewhere. Intercorrelations among number-correct scores, however, were available from Ree, Mullins, Mathews, and Massey (1982) for each of ASVABs 8a through 10b.

The median correlation coefficient across the ASVAB forms was determined for each pair of subtests. Coefficients from ASVAB Form 8b were most frequently the median. Therefore, the correlation matrix among the AFQT subtests for ASVAB Form 8b was chosen as most representative. The reported reliability coefficients for Form 8b from Ree et al. (1982) were used to correct these correlations for unreliability. The corrected correlation matrix was used as the true-score correlation matrix. The NO subtest was speeded and no reliability coefficient was reported. Therefore, the median correlation across all subtests was used as an estimate of the reliability for NO.

Sample Sizes and Combinations

As just described, 2,400 examinees were simulated for each of Samples X, Y, and Z. A subset of 1,000 examinees was randomly selected from each of the larger groups and constituted the smaller samples. Sixteen hundred examinees were selected from each of Samples X and Y on the basis of a separate selection composite and constituted two selected high-ability samples. These sample combinations were detailed in Table 4 above.

Generation of the True-Ability Distributions

Each examinee's true abilities for the subtests were sampled from the appropriate multivariate nonnormal distribution according to the procedure described in Vale and Maurelli (1983). This procedure is the multivariate extension of Fleishman's (1978) method for simulating nonnormal distributions. In this procedure, the target correlation matrix and marginal mean, variance, skewness, and kurtosis for each variable are specified in advance. The correlation matrix is then modified (see Vale & Maurelli, 1983, for details) and subjected to principal-components factorization. For each examinee, a (normally distributed) random number (i.e., component score) is generated for each component. The sum of the products of a variable's component loadings and the corresponding component scores defines an examinee's score on a variable. Fleishman's procedure is then applied separately to each variable score to yield a vector of variable scores for each examinee. Vale and Maurelli have shown that these score vectors have, asymptotically, the appropriate intercorrelations and marginal moments.

Throughout this project, normally distributed random numbers were generated by applying the Box-Muller transformation (Box & Muller, 1958) to random numbers uniformly distributed on the unit interval. All uniformly distributed random numbers were generated using a triple multiplicative congruential algorithm (Wichmann & Hill, 1982).

Table 7 presents the summary statistics of the distributions for Samples X, Y, and Z (and a separate evaluation sample, W) obtained after application of the Vale-Maurelli procedure. Sample W (described

Table 7
Summary Statistics of Multivariate Distributions of True Abilities: Samples X, Y, Z, and W

					Correlation coefficients			
Subtest	Mean	Variance	Skewness	Kurtosis	PC	AR	WK	NO
Sample X								
PC	0.050	0.761	0.128	-0.719	-			
AR	0.044	0.753	0.075	-0.666	.819	-		
WK	0.053	0.818	0.166	-0.875	.934	.791	-	
NO	0.681	0.038	-0.522	-0.294	.623	.687	.590	-
Sample Y								
PC	0.036	0.756	0.145	-0.710	-			
AR	0.022	0.750	0.141	-0.594	.820	-		
WK	0.040	0.824	0.169	-0.868	.936	.790	-	
NO	0.683	0.039	-0.550	-0.265	.625	.686	.597	-
Sample Z								
PC	0.208	0.678	0.201	-0.574	-			
AR	0.208	0.679	0.188	-0.459	.807	-		
WK	0.208	0.756	0.245	-0.757	.931	.765	-	
NO	0.702	0.034	-0.562	-0.167	.559	.624	.527	-
Sample W								
PC	0.038	0.746	0.134	-0.666	-			
AR	0.036	0.736	0.119	-0.584	.818	-		
WK	0.030	0.800	0.183	-0.811	.936	.787	-	
NO	0.681	0.038	-0.547	-0.214	.618	.671	.583	-

in more detail below) is composed of 10,000 examinees drawn from the parent (current) ability population and was used to evaluate all the equating transformations.

Comparing Tables 5 and 7 reveals that nearly all of the observed moments are slightly lower than those specified. However, the differences between the observed and specified moments are small and the two correlation matrices are similar enough to justify use of the procedure in this simulation.

Test Characteristics

Power Subtests

Test lengths. The current test lengths for subtests PC, AR, and WK are 15, 30, and 35 items, respectively. For this project, two different test lengths were simulated for each power subtest: 15 items and 30 items. These test lengths were chosen to model the test lengths of current subtests and to provide an effective test-length manipulation.

Specification of the true-item-parameter distributions. Distributions of the true item parameters were modeled after those obtained from items calibrated at the Navy Personnel Research and Development Center in San Diego. These data were provided by J. B. Sympson (personal communication, September 20, 1982) and were described by Sympson (1982). That paper described how items from ASVAB Forms 8, 9, and 10 were calibrated together with new prototype CAT items using LOGIST (Wood et al., 1976). Sympson provided item parameters for 90 PC items, 180 AR items, and 210 WK items. Table 8 presents the summary statistics for these three sets of item parameters. These statistics were used to specify the multivariate distributions of true item parameters for the three power subtests.

The correlations among the estimated item parameters obtained from Sympson were used to specify the correlations among true item parameters needed for the Vale-Maurelli procedure. These correlations were also reported in Table 8.

Generation of the true-item-parameter distributions. An entire pool of items was first generated for each subtest. Items were then assigned to individual subtest forms in a manner that ensured parallelism across forms. This item-assignment strategy was used so as to model the manner in which test forms are actually constructed.

Table 8
Summary Statistics Used to Specify Multivariate Distributions of True Item Parameters: Subtests PC, AR, and WK

Parameter	Mean	Variance	Skewness	Kurtosis	Correlation coefficients		
					a	*b*	*c*
PC							
a	0.966	0.158	0.633	-0.053	-		
b	-0.446	0.969	0.218	0.954	.325	-	
c	0.233	0.002	0.305	0.674	.150	-.007	-
AR							
a	1.595	0.438	0.516	-0.234	-		
b	0.060	1.045	-2.533	9.447	.649	-	
c	0.228	0.004	0.118	-0.589	.136	.132	-
WK							
a	1.548	0.442	0.633	-0.278	-		
b	-0.385	0.972	-0.934	0.560	.603	-	
c	0.260	0.003	-0.186	-0.424	.059	-.025	-

Note. Data from which these statistics were obtained were provided by J. B. Sympson (personal communication, September 20, 1982).

The true item parameters were generated for each power subtest using the Vale-Maurelli procedure and the following restrictions:

(a) $0.4 \le a \le 2.5$;
(b) $-3.0 \le b \le 3.0$; and
(c) $0.0 \le c \le 0.5$.

Items that fell outside these bounds were discarded and replaced.

Each short test contained 15 items and each long test contained 30 items. There were four short forms and four long forms for each subtest (see Table X1). Thus, each subtest required 180 items. In addition, the anchor tests required 90 items (two 15-item forms for each of three difficulty levels). Also, 30 selection-test items were generated. Thus a total of 300 items were required for each subtest.

Once item parameters were generated, they were modified to simulate tests of different difficulties. One hundred twenty items were made more difficult by adding a constant of 1.0 to the *b* parameters; *a* and *c* remained unchanged. Thirty items were used to construct anchor tests of medium difficulty; these items were modified by adding a constant of 0.50 to *b*. The remaining items were called "easy" items and were not modified at all. In all cases, if the resulting *b* parameter was greater than 3.0, all the parameters for that item were discarded, and a new set of item parameters was selected and modified accordingly. This process was repeated until there were enough items and the restrictions on all the parameters were met.

Tables 9 through 11 present the summary statistics for Subtests PC, AR, and WK, respectively, by difficulty level and overall. Discrepancies between the observed and targeted item parameters were small. The mean discrimination parameters for the three subtests were targeted to be 0.966, 1.595, and 1.548, respectively. Observed mean discriminations were 1.001, 1.459, and 1.430, respectively. The *c* parameters varied little about their targeted values. The *b* parameters were explicitly varied; the unmodified parameters, however, were close to their targeted values.

Differences in the mean discrimination parameter across difficulty levels can be observed, however. This is readily apparent for PC where the medium-difficulty items had a mean *a* parameter of

Table 9
Summary Statistics of Multivariate Distributions of True Item Parameters: Subtest PC

					Correlation coefficients		
Parameter	Mean	Variance	Skewness	Kurtosis	a	b	c
Easy (*n*=135)							
a	1.019	0.178	0.859	0.154	-		
b	-0.433	0.976	0.145	-.096	.219	-	
c	0.240	0.003	0.670	0.793	.226	.059	-
Medium (*n*=30)							
a	0.881	0.097	0.307	-1.042	-		
b	0.273	1.052	0.767	0.229	.215	-	
c	0.241	0.003	0.976	1.415	.098	-.060	-
Difficult (*n*=120)							
a	1.011	0.166	0.621	-0.398	-		
b	0.413	1.013	-0.072	-0.300	.352	-	
c	0.235	0.002	0.795	3.199	.035	.003	-
Overall (*n*=285)							
a	1.001	0.166	0.767	0.005	-		
b	-0.002	1.168	0.113	-0.201	.240	-	
c	0.238	0.002	0.795	1.674	.141	.006	-

Note. *n* is the number of items.

Table 10
Summary Statistics of Multivariate Distributions of True Item Parameters: Subtest AR

					Correlation coefficients		
Parameter	Mean	Variance	Skewness	Kurtosis	a	b	c
Easy (*n*=150)							
a	1.427	0.262	0.156	-0.973	-		
b	0.019	0.818	-1.373	1.235	.835	-	
c	0.222	0.004	0.236	0.094	.180	.240	-
Medium (*n*=30)							
a	1.600	0.344	-0.128	-1.379	-		
b	0.687	0.586	-1.145	-0.074	.881	-	
c	0.225	0.003	1.016	0.727	.052	-.032	-
Difficult (*n*=120)							
a	1.465	0.276	-0.026	-1.064	-		
b	1.096	0.820	-1.369	0.831	.843	-	
c	0.222	0.005	0.303	-0.409	.120	.139	-
Overall (*n*=300)							
a	1.459	0.278	0.069	-1.065	-		
b	0.517	1.056	-0.884	0.525	.747	-	
c	0.222	0.004	0.313	-0.071	.141	.151	-

Note. *n* is the number of items.

0.881 and the other items had a mean *a* parameter greater than or equal to 1.011.

Differences in the mean *a* parameters across difficulty levels for the AR items were also apparent. The easy items, for example, had a mean discrimination parameter of 1.427, compared to 1.465 and 1.600 for the difficult and medium items, respectively. Mean discriminations for the WK items ranged from 1.375 to 1.533 for the easy and medium items, respectively.

Assignment of items to individual test forms. Items for each subtest were assigned to the individual test forms, anchor tests, and selection tests in a manner that ensured parallelism across test forms. First, the items for each subtest were separated into the three different difficulty levels (easy, medium, or difficult, depending on the constant added to the *b*s); the item-assignment procedure was performed separately for each level.

All items at a specific difficulty level for a subtest were first sorted into a test-form-by-stratum matrix (see Table 12). That is, items were sorted according to the *b* parameter and assigned to 15

Table 11
Summary Statistics of Multivariate Distributions of True Item Parameters: Subtest WK

					Correlation coefficients		
Parameter	Mean	Variance	Skewness	Kurtosis	a	b	c
Easy (*n*=150)							
a	1.375	0.315	0.360	-1.193	-		
b	-0.516	0.796	-0.545	-0.545	.619	-	
c	0.256	0.004	-0.151	-0.519	.104	-.095	-
Medium (*n*=30)							
a	1.533	0.226	-0.139	-1.126	-		
b	0.063	0.371	-0.254	-0.631	.220	-	
c	0.262	0.003	-0.304	-0.431	.180	-.014	-
Difficult (*n*=120)							
a	1.474	0.209	0.140	-0.574	-		
b	0.678	0.673	-0.834	0.525	.545	-	
c	0.268	0.003	-0.366	-0.243	-.074	-.010	-
Overall (*n*=300)							
a	1.430	0.267	0.199	-1.034	-		
b	0.020	1.021	-0.435	-0.240	.517	-	
c	0.261	0.004	-0.257	-0.424	.057	.003	-

Note. *n* is the number of items.

strata so that Stratum 1 contained the items with the highest *b* values and Stratum 15 contained the items with the lowest *b*s. The number of test forms varied for the easy, medium, and difficult tests. Easy items, for example, were sorted into a 9- (for PC which had a shorter selection test) or 10- (for AR, WK) by-15 matrix. Each 30-item test form was constructed from two parallel 15-item tests. Two different 15-item anchor tests were constructed. (Only one of these forms was ever used for actual test equating; the second anchor test was constructed so that parallel-forms reliability could be computed). In addition, easy items were assigned to a 15-item (for PC) or 30-item (for AR, WK) selection test. Thus, the easy items were assigned to 9 or 10 different test forms.

Medium-difficulty items were required only for the anchor tests and, therefore, were assigned only to two different test forms. Similarly, difficult items were assigned to 8 test forms and 15 strata.

For each difficulty level, the items in the first stratum were permuted; i.e., each item was assigned to a test form at random. The items in the subsequent strata were assigned to test forms such that the

Table 12
Strategy for Assigning Items to Individual Subtest Forms

	Stratum				
	Difficult	-------->			Easy
Test forms	1	2	...	14	15
Easy					
anchor 1	X	X		X	X
anchor 2	X	X		X	X
form 1	X	X		X	X
form 2	X	X		X	X
form 3	X	X	...	X	X
form 3	X	X		X	X
form 4	X	X		X	X
form 4	X	X		X	X
selection	X	X		X	X
selection (AR, WK only)	X	X		X	X
Medium					
anchor 1	X	X	...	X	X
anchor 2	X	X		X	X
Difficult					
anchor 1	X	X		X	X
anchor 2	X	X		X	X
form 5	X	X		X	X
form 6	X	X	...	X	X
form 7	X	X		X	X
form 7	X	X		X	X
form 8	X	X		X	X
form 8	X	X		X	X

mean discrimination across all test forms was equalized as much as possible. This was accomplished, stratum by stratum, by (a) computing the mean discrimination for the items assigned to a test form so far, (b) computing the deviation of a test's current mean discrimination from the (grand) mean over all the items, and (c) assigning items to test forms within the current stratum such that the lowest-discriminating item was assigned to the test form with the largest positive deviation from the grand mean. The last step was repeated until each item within the stratum was assigned a test form; this entire process continued sequentially for each stratum until items from all 15 strata were assigned. Tables 13, 14, and 15 present the results of this item-assignment strategy.

The item-assignment strategy created subtest forms with approximately equal mean item discrimination. Mean discriminations

Table 13
True Item Parameter Means for Each Test Form: Subtest PC

Test form	n	a	b	c
Subtest forms				
1	15	1.050	-0.438	0.275
2	15	1.018	-0.435	0.239
3	30	1.017	-0.410	0.227
4	30	1.011	-0.455	0.236
5	15	1.018	0.431	0.228
6	15	0.999	0.370	0.260
7	30	1.002	0.393	0.235
8	30	1.022	0.424	0.231
Anchor tests				
easy 1	15	1.010	-0.409	0.239
easy 2	15	1.019	-0.421	0.222
medium 1	15	0.889	0.283	0.227
medium 2	15	0.873	0.263	0.255
difficult 1	15	1.020	0.428	0.233
difficult 2	15	1.000	0.443	0.225
Selection test	15	1.016	-0.461	0.257
Overall	285	1.001	-0.002	0.238

varied little across test forms within difficulty level. Differences among mean b and c parameters (within difficulty level) were small.

Speeded Subtests

The time limit for the 50-item NO subtest in ASVAB Forms 8, 9, and 10 was simulated in this study by modeling the distribution of the number of items attempted by current examinees. In addition, a shorter test (with the administration time cut in half) was also simulated by assuming that the number of items attempted by each examinee was cut in half. Item difficulty was not explicitly varied.

Composites

Two different kinds of composite scores were defined. An AFQT composite was formed by unit weighting the number-correct score on each of the three power subtests and weighting the number-correct score from the speeded subtest by one-half. The sum of these weighted scores formed a composite score analogous to the AFQT. In addition, a

Table 14
True Item Parameter Means for Each Test Form: Subtest AR

Test form	n	a	b	c
Subtest forms				
1	15	1.429	0.000	0.222
2	15	1.421	0.001	0.226
3	30	1.427	0.052	0.232
4	30	1.424	0.001	0.227
5	15	1.463	1.094	0.213
6	15	1.464	1.098	0.229
7	30	1.463	1.098	0.229
8	30	1.464	1.081	0.202
Anchor tests				
easy 1	15	1.432	0.029	0.198
easy 2	15	1.420	-0.017	0.219
medium 1	15	1.591	0.691	0.231
medium 2	15	1.609	0.682	0.219
difficult 1	15	1.463	1.064	0.231
difficult 2	15	1.473	1.149	0.239
Selection test	30	1.431	0.036	0.220
Overall	300	1.459	0.517	0.222

power composite was formed by unit weighting and summing the number-correct scores from the three power subtests.

The characteristics of each composite were defined by the characteristics of its component subtests. That is, composite Form 1 (see Table 2) was constructed by appropriately weighting and summing the scores from Form 1 (i.e., easy, short) subtests. Similarly, Form 8 was constructed by appropriately weighting difficult, long subtests. Test length and difficulty remained constant across subtests within a composite, although they did vary across the composites being equated. Hence, the test-form characteristics and pairings presented in Tables 2 and 3 are applicable to composite scores as well as to the individual subtests.

Selection Composite

All items for the three power selection subtests were drawn from the pool of easy items. The assignment of items to these selection tests was described earlier; mean *a*, *b*, and *c* parameters for these tests were presented in Tables 13 through 15. The speeded

Table 15
True Item Parameter Means for Each Test Form: Subtest WK

Test form	n	a	b	c
Subtest forms				
1	15	1.389	-0.538	0.256
2	15	1.377	-0.529	0.269
3	30	1.374	-0.504	0.264
4	30	1.375	-0.518	0.253
5	15	1.476	0.714	0.301
6	15	1.484	0.667	0.270
7	30	1.468	0.668	0.277
8	30	1.468	0.666	0.262
Anchor tests				
easy 1	15	1.366	-0.545	0.240
easy 2	15	1.370	-0.491	0.252
medium 1	15	1.503	0.080	0.264
medium 2	15	1.562	0.047	0.260
difficult 1	15	1.478	0.673	0.242
difficult 2	15	1.480	0.701	0.253
Selection test	30	1.374	-0.504	0.255
Overall	300	1.430	0.020	0.261

selection subtest used the same matrix for response generation that was constructed earlier. The selection test for Subtest PC was 15 items long; Subtests AR and WK each contained 30 items. The speeded subtest contained 50 items.

A selection composite score was computed by weighting the number-correct scores on the three power tests by one and weighting the number-correct score on the speeded test by one-half. The weighted scores were then summed to form an AFQT-like composite score. The 1,600 highest-scoring examinees (i.e., the top two-thirds) were selected from each of Samples X and Y and constituted the "selected" samples.

Anchor Tests

For equating power tests, anchor-test difficulty was matched to the difficulty of the two tests being equated. That is, when an easy test was equated to an easy test, an easy anchor test was used. Similarly, a difficult test was equated to another difficult test through a difficult anchor test. When an easy test was equated to a difficult test, however, an anchor test of medium difficulty was used. These anchor-test specifications were presented in Table 3.

All power anchor tests were 15 items long. Composites of power subtests were directly equated to each other using 15-item anchor tests that were constructed from the first five items from each of the three subtest anchors.

The anchor test used for equating two speeded tests was an external, "separately timed" test. This anchor test was simulated by assuming that the number of items attempted by each examinee was equal to the number of items that examinee attempted on the short test. That is, the anchor test was "administered" with the time limit equal to the time limit of the short speeded test. This manipulation is analogous to the requirement that all anchor tests used for equating power tests were 15 items long, the length of the short power tests.

Generation of Item Responses

Prior to generating item responses to each subtest, a vector of true abilities was drawn for each examinee from the specified multivariate distribution. For the power subtests, these abilities were true theta values. The speeded-subtest abilities were true proportion-correct scores.

Power Subtests

For the power subtests, the true ability and item parameters were used to compute the probability of a correct item response using the three-parameter logistic IRT model. This probability value was compared to a random number uniformly distributed on the unit interval. If the random number was less than the probability of a correct response, the simulated examinee was said to have correctly answered that item. Otherwise, the examinee was said to have responded incorrectly (see, e.g., Ree, 1981). Successive applications of this algorithm yielded a vector of observed scored responses for each examinee. Sets of response vectors were generated for each combination of subtest, anchor test, selection test, and sample as required by the project design. Item scores were summed to form raw number-correct scores.

Raw number-correct scores were used to equate tests using conventional and strong true-score methods. The item responses were used in IRT and STST equating.

Speeded Subtests

For the speeded subtests, observed number-correct scores were generated for each examinee according to the binomial error model proposed by Pieters and van der Ven (1982). In this model, the probability of each number-correct score, conditional on the number of items attempted, is given by

$$P(R_i = r \mid A_i = a) = \binom{a}{r} \pi_i^r (1 - \pi_i)^{a-r} \qquad [12]$$

where R_i is the number-correct score for examinee i;

A_i is the number of items attempted by examinee i;

π_i is the correct-response probability for examinee i; and

lower-case letters denote specific values of the random variables.

The probability of a correct response to an item is assumed to vary across examinees but to remain constant across all items in the test for a given examinee. This assumption is called the constancy hypothesis and implies that more difficult items require longer response times. That is, it is assumed that an examinee's response time varies with each item so that the probability of a correct response remains constant for that examinee over all items. It is clear that an examinee's true number-correct score is the product of his or her precision (true proportion correct) and speed (number of items attempted). That is, each pair of precision and speed values yields a single true number-correct score. Different combinations of precision and speed, however, may yield the same true score on a speeded test. Thus, for any true number- or proportion-correct score sampled from the multivariate ability distribution, there may be several corresponding pairs of precision and speed values.

To generate item responses, an examinee's true proportion-correct score was first sampled from the appropriate multivariate distribution. This proportion-correct score was converted to a true number-correct score by multiplying by the number of items. This number-correct score was then compared to the distribution of the number of items attempted conditional on number correct. The number of items attempted by that examinee was randomly chosen from the number-of-items-attempted values corresponding to the specified number correct (weighting each cell by its proportion of cases). The true number-correct score was divided by the number of items attempted in order to calculate precision.

Individual item responses were then generated for the examinee by comparing the precision level to a random number uniformly distributed on the unit interval as described above for power subtests. The number of item responses generated for an examinee was equal to the number of items attempted. If the number of items attempted was less than the length of the subtest, the "not reached" items were scored as incorrect responses. The observed number-correct score for an examinee was the simple sum of the scored item responses. Raw scores were used to equate tests using conventional and strong true-score methods; item responses were also used for STST. IRT was not applied to speeded tests.

Adequacy of the Simulation Procedures

A faithful simulation procedure should produce simulated observed test scores similar to test scores actually obtained by ASVAB examinees. Accordingly, the summary statistics for ASVAB 8b (reported in Ree et al., 1982, Tables 3 and 18) were used as a basis for comparison with the simulated test scores. Examinee responses to the Form 3 (easy, long) AR, WK, and NO subtests and to the Form 1 (easy, short) PC subtest were used to maximize test-form comparability between the real and simulated data sets.

The summary statistics used for this comparison are presented in Table 16. The mean subtest scores for the two data sets were very similar, in general less than two raw-score points; the lone exception was for the Word Knowledge subtest. However, the simulated WK subtest contained 30 items, whereas the real WK subtest contained 35 items. When the mean score for the simulated test is converted to a proportion correct and then multiplied by 35, the resulting figure is 24.91, less than half a score point different from the real data. The real-data variances were uniformly larger than those from the simulated data; the higher-order moments for the two data sets were less dramatically -- and less consistently -- different. In general, the real-data correlations were larger than those computed from the simulated data. The rank order of the correlations, however, was

Table 16
Summary Statistics for Real and Simulated ASVAB Subtest Scores

						Correlation coefficients			
Subtest	n	Mean	Variance	Skewness	Kurtosis	PC	AR	WK	NO
Real data (N=2,510)									
PC	15	10.33	11.49	-0.65	-0.41	-			
AR	30	18.52	54.91	-0.11	-1.10	.71	-		
WK	35	24.60	59.91	-0.69	-0.41	.81	.73	-	
NO	50	35.77	102.82	-0.63	-0.01	.55	.64	.56	-
Simulated data (N=2,400)									
PC	15	10.70	6.23	-0.41	-0.43	-			
AR	30	17.52	41.47	0.33	-0.93	.61	-		
WK	30	21.35	34.84	-0.18	-1.15	.69	.69	-	
NO	50	34.05	98.21	-0.52	-0.27	.52	.61	.55	-

Note. Real-data statistics were taken from Tables 3 and 18 (for ASVAB-8b) in Ree et al. (1982). N is the number of examinees.

virtually the same for the two sets; it is likely, therefore, that the differences in the levels of the correlations reflected the differences in the variances of the two data sets.

Applications of Equating Transformations

Linear Equating

The data needed to linearly equate two tests are score means and standard deviations. In this study, the equated score (X'_{old}) was directly obtained for the single-group and equivalent-group designs by the equation

$$X'_{old} = (X_{new} - \overline{X}_{new})(sd_{old}/sd_{new}) + \overline{X}_{old} \qquad [13]$$

For anchor-test equating, each test was separately equated (in each group) to the anchor test by Equation 13. Scores on the two tests that were equated to the same score on the anchor test were considered equated to one another.

Linear interpolation was applied as needed to equate each new-test score to the old-test score having the same equated anchor-test score (between zero and the maximum score). Linear extrapolation was used to complete the equating table for unequated high and low score values on the new test, as necessary. Unequated low scores were defined as those that had an equated anchor-test score that was less than zero or below the lowest anchor-test score that was equated to any score on the old test. The extrapolation line for these scores was the extension of the line connecting the lowest equated old-test score and a point one third of the way toward the highest equated score. An analogous procedure was followed for unequated scores at the high end of the new test.

Equated scores that fell outside the range delimited by zero and the maximum score on the old test were set equal to the nearer endpoint. No corrections were made for unequal reliabilities.

Equipercentile Equating

Equipercentile equating is done in a series of steps. First, raw percentile tables are computed, and corresponding raw scores are set equal. In addition, percentile tables and/or the equating table can be smoothed; the two smoothing steps are optional. In this study, five variations of equipercentile equating were examined: (a) no smoothing was performed at all; (b) the equating table was smoothed using cubic polynomial regression, and percentile tables were not smoothed; (c) the equating table was smoothed using cubic splines, and the percentile tables were not smoothed; (d) percentile tables were smoothed using cubic polynomial regression, and the equating table was not smoothed; and (e) percentile tables were smoothed using cubic splines, and the equating table was not smoothed. Equating and smoothing were always performed using real-valued raw scores; equated scores were rounded to integers only at the very last step (i.e., at the evaluation phase).

Components of equipercentile equating are described in detail because different implementations are possible. A subset of the smoothing procedures was selected and applied throughout the study; the data used to choose among these smoothings for the main study are presented below.

Components of the Equating Procedure

Percentile tables. The raw frequency distribution of total scores on a test was obtained and transformed to a percentile distribution. The percentile rank for a score was computed on the score midpoint (i.e., all the cases below a score plus half the cases at the score).

Regression smoothing of percentile tables. Percentiles for a single test (old, new, or anchor) were regressed on corresponding test scores using cubic polynomial regression; only those test scores with reliable data (i.e., those having observed percentiles within the 0.1-99.9 range) were included in this regression. The resulting regression weights were applied to the same scores to obtain smoothed percentiles. If the smoothed percentile was less than 0, greater than 100, or nonmonotonic (rising for lower scores, declining for higher scores), the corresponding score was removed from the smoothed table and later replaced by an extrapolated value (see below). Only the smoothed portion of the percentile table was used in the initial equating phase; the tails of the equating table were extrapolated later in the equating procedure.

Spline smoothing of percentile tables. Reinsch's (1967) cubic-splines algorithm was used to smooth the percentile table for an individual test. A moderate smoothing parameter value (one-half the number of scores values), as suggested by Kolen (1983), was used to control the degree of smoothing. Each score point was weighted by its standard error (Guilford, 1965, p. 161):

$$se_i = \sqrt{p_i(1-p_i)/N} \qquad [14]$$

where se_i is the standard-error weight applied to score i and p_i is the percentile rank of score i. Again, only the smoothed portion of the percentile table, as defined in the previous section, was used in the initial equating phase.

Equating procedure. For single-group and equivalent-groups equating, a score on the new test was equated to the score on the old test having the same percentile. If the percentile on the new test fell between percentiles for two scores on the old test, one of the following procedures was used. For unsmoothed percentile tables, linear interpolation was used to obtain the equated score for scores on the new test having percentiles within the (0.1 to 99.9) range and between the percentile values for the old test's lowest and highest scores. For regression-smoothed percentile tables, the regression curve for the old test was used to interpolate between old-test points using a Newton-Raphson iterative solution of the third-degree polynomial. For spline-smoothed percentile tables, the appropriate spline equation was used to interpolate between each pair of old-test points, again using Newton-Raphson methods.

For unsmoothed equating tables (whether or not the percentile tables were smoothed), linear extrapolation was applied to obtain equated scores in the tails of the table, as needed. Otherwise, the equating table was smoothed (as described below); if necessary, extrapolation to the tails of the table was performed after smoothing.

For anchor-test equating, each of the tests to be equated was first separately equated (within each group) to the common anchor test by the single-group equipercentile equating method. The linear interpolation and extrapolation methods described above for linear anchor-test equating were used to equate a score on the new test to the score on the old test having the same equated anchor-test score.

Regression smoothing of equating tables. Using pairs of scores, equated old-test scores were regressed on corresponding new-test scores using cubic polynomial regression. Linear extrapolation was performed as described previously to obtain equated scores for new-test scores not previously equated to the old test (i.e., outside the range of reliable data) or that occurred beyond a point of inflection in the upper or lower tail. The resulting smoothed equating transformation for the new test was bounded by zero and the maximum score on the old test (i.e., equated scores outside this range were set equal to the specified bound).

Spline smoothing of equating tables. Reinsch's (1967) cubic-splines algorithm with a moderate value (cf. Kolen, 1983) for the smoothing parameter was used to smooth the obtained equating tables. The standard errors of equipercentile equating (adapted from Kolen, 1983, p. 7) were used to weight the individual score points:

$$se_i = \frac{1}{k}\sqrt{\{[mp_i(1-p_i)/g^2] - [(p_i - p_{old/less}) \cdot (p_{old/more} - p_i)/(N_{old})(g^3)]\}} \qquad [15]$$

where se_i = standard-error weight applied to score i;

k = the number of items on the old test;

$100 * p_i$ = percentile rank for score i on the new test;

$p_{old/less}$ = largest percentile on the old test $< p$;

$p_{old/more}$ = smallest percentile on the old test $\geq p$;

$g = p_{old/more} - p_{old/less}$; and

$$m = \frac{1}{N_{old}} + \frac{1}{N_{new}} .$$

Note that Kolen presented the standard errors for equated integer scores that range from 0 to k. His equation was modified here (i.e., multiplied by 1/k) to account for the fact that all equating tables in this study were presented in a proportion-correct metric. The equating transformation was completed by linear extrapolation, as necessary, and bounded as described in the previous section.

Comparison of the Smoothing Procedures

All five smoothing methods were applied to the equipercentile equating of the AR subtest. True-score-based RMSE and bias were calculated for each application of equipercentile smoothing. (These error indices are described in more detail below.) A tally was taken of the "best" smoothing method (i.e., that having the lowest error) across equatings for each error index; this tally is presented in Table 17. This tally indicated that the regression methods performed somewhat better than did the spline methods. Summary error indices for the smoothing methods are also presented in Table 17. According to this criterion, there was little difference among the methods, except that the RMSE was slightly higher for the regression smoothing of percentile tables. None of the smoothing methods outperformed "no smoothing."

Three smoothing procedures were performed on all remaining tests and composites: no smoothing anywhere, regression smoothing of equating tables, and regression smoothing of percentile tables. These smoothing procedures were selected to provide a good comparison of methods that are widely used in practice (equating-table smoothing), seem more appropriate theoretically (percentile-table smoothing), and are supported by preliminary data-analysis results (no smoothing).

Item Response Theory Equating

The data required for IRT equating are the linked item parameter estimates for the two tests to be equated. For both the single-group and anchor-test designs, item calibration was applied to a single matrix containing item responses from both tests; because all items were simultaneously calibrated, no additional linking was necessary.

Table 17
Equipercentile Equating Smoothing Methods: True-Score Error Indices and Tally of "Best" Method for Subtest AR

	Parallel		Nonparallel		"Best" tally	
Smoothing method	RMSE	Bias	RMSE	Bias	RMSE	Bias
Unsmoothed	0.008	0.002	0.045	0.008	23	22
Regression-smoothed						
Percentile tables	0.011	0.001	0.053	0.009	20	30
Equating table	0.008	0.002	0.046	0.011	16	8
Spline-smoothed						
Percentile tables	0.008	0.002	0.047	0.010	9	5
Equating table	0.007	0.002	0.047	0.011	13	16
N of Equatings	36		45		81	

For the single-group design, all examinees responded to all items. For the anchor-test design, the examinee-by-item response matrix included data that was coded as "not reached" for the test not administered to an examinee sample. For the equivalent-groups design, each test was calibrated separately. The assumption of equivalent groups implies that the two sets of item parameter estimates were automatically linked; no additional linking procedure was implemented.

Item Calibration Program

IRT parameters were computed using the program ASCAL (Assessment Systems Corporation, 1982). ASCAL is a conditional maximum-likelihood/modal-Bayesian item calibration program for the three-parameter logistic item response model. The maximum likelihood algorithms are similar to those presented by Wood et al. (1976) and used in the program LOGIST. However, ASCAL differs from LOGIST in the following ways.

In ASCAL, Bayesian priors have been added to the ability estimates and to the *a* and *c* parameters. A standard normal distribution is used for ability. For the *a* parameter, a Beta distribution is used with both shape parameters equal to 3.0 and endpoints equal to 0.3 and 2.6. For the *c* parameter, a Beta distribution is used with shape parameters equal to 5.0 and endpoints equal to -0.05 and (2/n)+0.05, where *n* is the number of alternatives.

The ability estimates are unbounded; the Bayesian prior distribution imposed on ability prevents the ability estimates from becoming infinitely large or small. The a parameter is bounded between 0.4 and 2.5, the b parameter is bounded between -3.0 and 3.0, and the c parameter is bounded between 0.0 and (2/n).

The estimation process begins with the computation of standardized number-correct scores for the examinees and conventional proportions correct and item-total biserial correlations for the items. These statistics are then transformed into IRT a and b parameters using Jensema's (1976) transformations; c parameters equal to (1/n) are assigned to the items in this initial stage.

These initial parameter estimates are then used to estimate abilities, and examinees are grouped into 20 fractiles, each fractile containing approximately five percent of the examinees. The fractile means are computed and standardized (i.e., the mean of the means is set to zero and the standard deviation of the means is set to one). Item parameters are then estimated using the fractile means and frequencies as input data.

The ability and item parameter estimation process is repeated until the parameter estimates converge or until ten iterations have been performed. If an estimate has not converged in ten iterations, the current value is used.

Equating Procedure

Equated number-correct scores correspond to the same theta. The theta that would result in a true score equal to a given new-test score was found, bounded by ± 4.5, and inserted into the true-score formula for the old test in order to obtain the equated score:

$$X'_{old} = \sum_{i=1}^{N_{old}} P_i(\theta) \qquad [16]$$

This was done for each new-test score between the chance true score and a perfect score, exclusive, on the new test. Linear extrapolation, as described above, was used to extend equating to the lower and upper ends of the equating table.

Strong True-Score Theory Equating

Strong true-score theory produces an estimated distribution of true scores from a sample distribution of observed scores. The true-score distributions are then equated such that a score on the new test is equated to the score on the old test having the same estimated true percentile.

Estimating a Test's True-Score Distribution

The general STST model (Lord, 1980, Equation 16-2) defines the relationship in the population between observed scores (x) and true scores (ζ) as

$$\phi(x) = \int_0^1 g(\zeta)\, h(x|\zeta)\, d\zeta \qquad [17]$$

where $\phi(x)$ is the population frequency distribution of observed scores;
$g(\zeta)$ is the true-score density at ζ;
$h(x|\zeta)$ is the conditional distribution of observed scores given true score;
$x = 0, 1, \ldots n$; and
n is the number of items in the test.

The sample frequencies, f(x), are only a rough approximation to the population observed-score distribution, $\phi(x)$. Thus, the scores are grouped into U intervals (see Appendix A) to reduce irregularities. The objective is to find a $g(\zeta)$ that will produce an exact fit to the population $\phi(x)$. Any one of several smooth solutions, all smooth solutions being very close to one another, will suffice. Smoothness is measured (Lord, 1980, Equation 16-4) by

$$\int_0^1 \frac{\{g(\zeta) - \gamma(\zeta)\}^2}{\gamma(\zeta)}\, d\zeta \qquad [18]$$

where $\gamma(\zeta)$ is some smooth density function, either $\gamma(\zeta) \equiv 1$ or $\gamma(\zeta) \propto \zeta(\zeta - 1)$ being satisfactory.

Lord (1980, Equation 16-9) has shown that the "smoothest" solution (i.e., that having the smallest smoothness measure) is:

$$g(\zeta) = \gamma(\zeta) \sum_{u=1}^{U} \lambda_u \sum_{X:u} h(X|\zeta) \qquad [19]$$

where λ_u is a parameter of the observed-score distribution $\phi(x)$.

The general model thus reduces (Lord, 1980, Equation 16-11) to

$$\phi(x) = \sum_{u=1}^{U} \lambda_u a_{xu} \quad \text{for } x = 0, \ldots, n \qquad [20]$$

where $a_{xu} = \sum_{X:u} \int_0^1 \gamma(\zeta)\, h(X|\zeta)\, h(x|\zeta)\, d\zeta$

The a_{xu}'s are constants to be computed from the data; computational formulas are derived in Appendix A. The λ_u's are parameters of ϕ to be estimated and then substituted back into Equation 20 to obtain estimates of the true-score distribution as part of the equating procedure.

Initial λ estimates. Substituting sample values into Equation 20 yields (Lord, 1969, Equation 39, corrected for a notational error)

$$f_u = \sum_{x:u} f(x) = \sum_{v=1}^{U} \hat{\lambda}_v \sum_{x:u} a_{xv} \qquad [21]$$

Since the f_u's and a_{xv}'s are known, letting

$$A_{uu} \equiv \sum_{x:u} a_{xu} \qquad [22]$$

initial λ's can be obtained by solving the matrix equation

$$\hat{\lambda} = fA^{-1} \qquad [23]$$

The $\hat{\lambda}$'s must then be rescaled (see Appendix A) to keep all $\hat{\lambda}_u \geq 0$. This restriction guarantees all $\hat{g}(\zeta) \geq 0$ for $0 \leq \zeta \leq 1$ (Lord, 1980, p. 241) which is necessary for an acceptable solution.

Refining the λ estimates. Maximum likelihood estimation procedures that simultaneously use all the sample frequencies are most efficient in refining the λ's. The set of λ_v's that maximizes the likelihood function (Lord, 1980, Equation 16-10)

$$L = \prod_{x=0}^{n} \{\phi(x)\}^{f(x)} \qquad [24]$$

for the set of observed f(x)'s is found by the steps described in Appendix A.

Equating the Tests

Given sample values and final parameter estimates computed above for each test separately, the estimated true percentile (i.e., the estimated proportion of examinees in the population who would score below a given true score) can be computed for any score on the new or old test from the integral

$$\int_0^t g(\zeta)\, d\zeta \qquad [25]$$

where $\underline{t}$ = true proportion correct on the test (see Appendix A for the procedure). Equated scores on the old and new tests have the same estimated true percentiles on the two tests.

The endpoints were fixed: scores of zero and n_{new} on the new test were equated to scores of zero and n_{old} on the old test. For each score from 1 to (n_{new} - 1) on the new test, first the estimated true percentile on the new test was obtained, and then the equated score (the score on the old test having the same estimated true percentile) was obtained by STST methods if possible.

Strong true-score theory does a poor job estimating the tails of the distribution when few or no observed data fall there. Hence, the area of the old test for which STST equating was possible was defined as that in which the estimated true percentiles were between 0 and 100 and were monotonically increasing and for which the observed percentiles fell between 0.1 and 99.9. If the estimated true percentile on the new test fell outside the range of good values of estimated true percentiles on the old test, no equated score was returned. Otherwise, an initial value for the equated score was found and then Newton-Raphson iterative procedures were used to refine the equated score, i.e., to make it a value whose estimated true percentile was actually equal to that of the new-test score within a certain tolerance. (Appendix A describes both these steps.)

Linear extrapolation was performed on the line joining zero and the lowest equated old-test score for unequated new-test scores in the lower tail; the line joining n_{old} to the highest equated old-test score was used for unequated scores in the upper tail.

Procedures for Equating Test Composites

The power and AFQT composites were equated in three different ways. First, the composite scores themselves were directly equated by applying the conventional equating transformations to the composite scores in exactly the same manner as was done for the scores on the individual subtests. This was done to equate power to power composites, AFQT to AFQT composites, and power to AFQT composites. Power composites were also directly equated using strong true-score theory. In addition, both power and AFQT composites of equated subtests were formed; no further equating transformation was applied to these composite scores. Finally, both power and AFQT composite scores were indirectly equated using score statistics and correlations from individual subtests. The specific procedures and the data-collection requirements are detailed below.

Equating Composite Scores Directly

When examinees take all the subtests in a battery, composite scores can be computed as the weighted sum of individual unequated subtest scores. The composite scores themselves can then be directly equated. Because the goal of composite-score equating is to define equivalent scores on two composites of subtests, this direct procedure is the preferred method of equating composite scores.

When composite scores are directly equated, only one transformation table needs to be constructed and used. Subtest scores can be weighted as usual and combined into composites; a single equating transformation is then applied to these sets of composite scores.

In this study, composite scores were directly equated using the conventional and STST transformations. Because IRT assumes that each test is unidimensional, it is not applicable for equating multidimensional composite scores directly. The power composites were equated directly using all of the data collection designs, with the exception that the anchor-test design was not used with strong true-score theory. This exception was made because of the practical difficulties involved in applying strong true-score theory to the anchor-test composite (which was composed of five items from each of the individual subtest anchors). AFQT composites were not directly equated using the anchor-test design because of its impracticality, as discussed above.

Forming Composites of Equated Subtests

When each group of examinees is administered only a single subtest, composite scores cannot be equated by the direct methods. The only way in which any type of equivalence can be made between the two sets of composite scores is by first equating the individual subtests. Composite scores can then be formed for future examinees who are administered all the subtests in the new battery by applying the appropriate composite weights to their equated subtest scores.

With this procedure, a separate transformation table needs to be constructed and applied for each subtest in the composite. However, each equating transformation can be computed after the administration of individual subtests; it is not necessary to administer any more than one subtest (either one or two forms) to an intact group of examinees in order to equate the composites. The primary disadvantage of forming composites from equated subtests is that the resulting equating transformation contains errors from three or four separate and independent equating transformations and, therefore, probably contains a greater amount of error than does the equating transformation obtained when composite scores are directly equated.

This procedure can be used with every data collection design and every testing model and transformation form, with the exception that IRT can be used to equate power composites but not AFQT composites. The data obtained previously from equating individual subtests were used to equate composite scores by this method.

Equating Composite Scores Indirectly Through the Subtests

Composite scores can also be equated indirectly using conventional linear procedures that take into account the original composite weights, subtest means and standard deviations, and the intercorrelations among the subtest scores. This procedure is a reformulation of the linear equating model in which two composite scores are considered to be equated if their corresponding standard scores are equal.

The formulae for performing this type of composite equating were derived as follows:

$$Y_O = [W_O'] \, [X_O] \quad + \quad C_O \tag{26}$$

$$Y_N = [W_N'] \, [X_N] \quad + \quad C_N \tag{27}$$

where Y_O and Y_N are, respectively, the old and new composite scores for an examinee;

$[X_O]$ and $[X_N]$ are vectors of old and new subtest scores;

$[W_O]$ and $[W_N]$ are vectors of old and new weights applied to the individual subtests to yield composite scores; and

C_O and C_N are the old and new constants applied to yield composite scores.

The equation for linearly equating composite scores Y_O and Y_N is

$$Y_O = aY_N + b \tag{28}$$

where

$$a = \sqrt{\frac{[W_O'][V_O][W_O]}{[W_N'][V_N][W_N]}} \; ;$$

$$b = [W_O'][\bar{X}_O] + C_O - a * ([W_N'][\bar{X}_N] + C_N) \; ;$$

$[V_0]$ and $[V_N]$ are the variance-covariance matrices of the old and new subtest scores; and

$[\bar{X}_0]$ and $[\bar{X}_N]$ are the mean vectors of the old and new subtests.

This procedure is equivalent to the linear procedure for equating composite scores directly when each group of examinees takes all the subtests in a battery.

An advantage of this indirect procedure is that it can be used to equate test batteries with partial data under certain circumstances. It can be applied when examinees do not take all the subtests in a battery. There are two requirements: (a) a subset of the examinees must have taken each possible pair of subtests so that the intersubtest correlations can be estimated for each battery, and (b) the distinct examinee subgroups must be randomly sampled from the same population. When examinees take only a subset of the subtests in any battery, the subtest statistics are computed from the responses of several distinct subgroups of examinees. These values can be used as estimates of those that would have been obtained if the entire battery had been administered to a single group of examinees. Under these conditions, this procedure is an approximation to the procedure for equating composite scores directly. Examination time can be reduced if the entire battery does not have to be administered to an intact group of examinees.

The responses from examinees who took only selected pairs of subtests were used to equate composite scores using the linear procedure. Both the single-group and equivalent-groups data collection designs were investigated.

Linear equating procedures were applied to partial data sets where examinees did not receive all the subtests in the battery. Two of the subtests were administered to each examinee subgroup in a manner that ensured that all possible test pairs were administered. The manner in which these subtest pairs were administered to the different examinee subgroups is presented in Table 18.

Power-test composites were composed of three different subtests. Thus, three distinct examinee subgroups were required to administer the three possible subtest pairs (Subtests 1 and 2, Subtests 1 and 3, and Subtests 2 and 3). Since subtest scores were available from 2,400 examinees in each sample, each subtest pair was administered to a distinct subgroup of 800 examinees. When equating transformations were based on subtest scores of selected examinees, sample sizes were correspondingly smaller. For each of the three power subtests, then, score data were available from 1,600 unselected and approximately 1,067 selected examinees.

Table 18
Administration of Subtests to Examinee Subgroups: Creating Partial Data Sets

Subtests administered	Sequence number of examinees in each subgroup	
	Unselected	Selected
Power-test composites		
1 and 2	1- 800	1- 533
1 and 3	801-1600	534-1067
2 and 3	1601-2400	1068-1600
AFQT composites		
1 and 2	1- 400	1- 267
1 and 3	401- 800	268- 533
2 and 3	801-1200	534- 800
1 and 4	1201-1600	801-1067
2 and 4	1601-2000	1068-1333
3 and 4	2001-2400	1334-1600

AFQT composites were composed of four distinct subtests. Thus, examinees were divided into six distinct subgroups and were administered one of the six possible subtest pairs. Each unselected subgroup contained 400 examinees; subgroups of selected examinees were two-thirds that size. For each of the four AFQT subtests, score data were available from 1,200 unselected and 800 selected examinees.

Evaluative Criteria

Error in an equating transformation was isolated and evaluated in this study by applying the transformation to true-score data from a separate "evaluation" sample of examinees. Sample W abilities for the four subtest areas were generated for 10,000 new examinees sampled from the parent population distribution (i.e., the multivariate distribution of abilities that defined Samples X and Y). These abilities were thetas for the three power subtests and proportions correct for the speeded subtest. This sampling approach to the generation of ability distributions was used instead of numerical integration over a density function because the density function does not exist for the nonnormal distributions sampled.

The thetas, in combination with the true item parameters and the three-parameter logistic IRT model, were used to generate true proportion-correct scores for this sample on every power test. The true number-correct scores sampled for the speeded tests were converted to true proportion-correct scores. Additionally, observed proportion-correct scores were generated for every examinee in the evaluation sample using a random-number process in conjunction with the IRT model for power tests and the binomial error model for speeded tests. True composite scores were obtained for each examinee by weighting the true scores on each subtest and summing across subtests. The evaluative indices described below were computed on the evaluation sample so that the sample size and composition remained constant for all equatings evaluated.

True proportion-correct scores on the new test were equated to proportion-correct scores on the old test by applying the equating transformations computed from observed response data. The difference between the equated old-test score and the true old-test score was then computed for each examinee in the evaluation sample; functions of these difference scores were calculated as global indices of equating accuracy. The specific indices that were computed included root mean squared error (RMSE) and bias. RMSE is the square root of the mean squared difference between the true and equated old-test scores. Bias is the difference between the mean true score and the mean equated score on the old test.

Real-Data Application

Raw Data

Item response data for the real-data application phase of this project were obtained from Task II of the Omnibus Item Pool and Test Construction Project (Prestwood, Vale, Massey, & Welsh, in press). The items were part of the initial operational item pool for the adaptive ASVAB and were administered to MEPS examinees during the calibration phase of the Omnibus project from May to July 1983. Both male and female examinees were included. During this phase of the Omnibus project, items were randomly assigned to specific test booklets. Within each booklet the items were randomly ordered.

All item response data were edited. A redundantly coded form number allowed improperly recorded booklet numbers to be detected and, in some cases, correct booklet numbers to be recovered. A patterning coefficient was developed to detect response patterns ("ABCABC") and response strings ("AAAA"). Examinees who exhibited response patterns and strings and who responded to fewer than six items were deleted from the data set. Less than 0.25% of the examinees were deleted during this process. For details concerning this data-editing process, see Prestwood, Vale, Massey, and Welsh (in press). Real data analyses for this project were based on item responses to a test booklet containing 86 Word Knowledge items.

In order to parallel the computer simulations as closely as possible, the following procedures were performed. First, three examinee groups were defined. The first 1,000 examinees from the Omnibus data file formed Group 1, the next 1,000 examinees formed Group 2, and the next 1,000 examinees formed Group 3. Groups 1 and 2 were used for equating; Group 3 was used as a hold-out evaluation sample.

Items were assigned in a counterbalanced order to two 30-item tests and a 15-item anchor test (only 75 of the 86 items were used). Items were assigned to each test in an "ABCBA" format, where "A" denotes assignment to the "old" test, "B" denotes assignment to the "new" test, and "C" denotes assignment to the anchor test. Examinee responses were scored and total test scores were computed for each examinee.

Data Collection Designs

The single-group, equivalent-groups, and anchor-test data collection designs were used to equate tests using real examinee data. Group 1 responses to the old and new tests were used to equate the test using the single-group design. The responses of Group 1 to the old test and the responses of Group 2 to the new test were used to equate the two test using equivalent groups. Similarly, responses of Group 1 to the old test and the responses of Group 2 to the new test were also used to equate tests using the anchor-test design; in addition, anchor-test responses for the two groups were used.

Equating Transformations

Linear, equipercentile, IRT, and STST procedures were used to equate the two sets of test scores. The applications of these equating transformations were identical to those described above in the simulation procedures.

Evaluative Criteria

The criterion for evaluating equating accuracy, using real data, differed somewhat from the criterion used in Monte Carlo simulations. When real data are used, an examinee's true scores are not known; only the observed scores on the two tests are available. The differences between the observed old-test scores and the equated old-test scores are a measure of how well the equating procedure can recover the scores actually obtained by the examinees. The standard error of the difference between the observed and equated scores was computed and served as the base for evaluating the observed-score RMSE and bias indices of equating accuracy.

For all three designs, the equated test scores can be compared to the scores actually obtained by the examinees in the evaluation sample, Group 3.

RESULTS AND DISCUSSION

Choosing an Equipercentile Smoothing Method

Results

Table 19 presents the results from the three smoothing methods applied to each case of equipercentile equating. The true-score error indices are presented for (a) all power subtests, (b) speeded subtests, and (c) all composites (except the indirect composites).

Table 19
True-Score Error Indices for Equipercentile Smoothing Methods

	Tests/Composites			
	Parallel		Nonparallel	
Smoothing method	RMSE	Bias	RMSE	Bias
Power subtests				
Unsmoothed	0.008	0.000	0.036	0.008
Smoothed percentile tables	0.009	-0.001	0.041	0.009
Smoothed equating tables	0.008	0.000	0.037	0.010
N of equatings	108		135	
Speeded subtests				
Unsmoothed	0.010	0.001	0.012	0.002
Smoothed percentile tables	0.036	-0.010	0.039	-0.010
Smoothed equating tables	0.015	0.004	0.017	0.004
N of equatings	36		45	
Composites				
Unsmoothed	0.018	0.002	0.032	0.009
Smoothed percentile tables	0.018	0.002	0.036	0.009
Smoothed equating tables	0.018	0.003	0.033	0.009
N of equatings	168		210	

Note. All smoothing procedures were based on cubic polynomial regression.

In general, the differences among the smoothing methods were small for the parallel subtests. For the parallel power subtests, the results across smoothing methods were virtually identical; the regression smoothing of the percentile tables was markedly poorer (according to both error indices) for speeded subtests. The RMSE for this case was 0.036; the corresponding RMSE values were 0.010 and

0.015 for the unsmoothed and regression-smoothed equating tables, respectively. Bias indices followed the same pattern as the RMSEs.

Error indices from the nonparallel-test equatings were larger than those from the parallel-test equatings, with RMSEs ranging from 0.012 for unsmoothed speeded tests to 0.041 for smoothed percentile tables. As before, regression-smoothed percentile tables resulted in larger errors than did any other smoothing method.

The error indices obtained when composites were equated were larger, in general, than those observed for individual subtests. As before, errors were larger for the nonparallel-composite pairings (RMSEs of 0.032-0.036) than for the parallel pairings (RMSEs of 0.018). All three smoothing conditions performed equally well for the parallel composites. For the nonparallel composites, regression smoothing of the percentile tables was slightly worse (in terms of RMSE) than the other two smoothing methods.

Discussion

When parallel power subtests and composites were equated, all three smoothing methods yielded comparable amounts of error; in all other cases, regression smoothing of percentile tables typically resulted in larger errors than did the other smoothing methods. Neither type of regression smoothing improved upon "no smoothing" for any condition; when differences were observed among the smoothing methods, they tended to favor "no smoothing." Hence, the remainder of the comparisons presented in this report are based only on the unsmoothed equipercentile equating tables.

Equating Individual Subtests

Equating Methods

Results

Table 20 reports the true-score error indices computed when parallel subtests were equated. As this table shows, there were only small differences among the equating methods when they were applied to parallel power subtests. The true-score RMSEs for IRT and STST methods were slightly larger (by 0.002-0.004 points) than those from the conventional methods; all methods were essentially unbiased. Linear equating outperformed equipercentile and STST methods when parallel speeded subtests were equated (RMSE of 0.004 vs. 0.010-0.015).

Table 20 also presents the true-score error indices computed when nonparallel subtests were equated. IRT and STST equatings were

Table 20
True-Score Error Indices for Equating Subtests

	Type of subtest			
	Power		Speeded	
Equating method	RMSE	Bias	RMSE	Bias
Parallel subtests				
Linear	0.007	0.000	0.004	-0.001
Equipercentile	0.008	0.000	0.010	0.001
IRT	0.010	0.000	-	-
STST	0.011	0.001	0.015	-0.002
_N_ of equatings	108		36	
Nonparallel subtests				
Linear	0.048	0.006	0.006	-0.001
Equipercentile	0.036	0.008	0.012	0.002
IRT	0.024	0.005	-	-
STST	0.021	-0.001	0.015	-0.002
_N_ of equatings	135		45	

clearly superior (in terms of RMSE) to the conventional methods for equating power subtests. The RMSEs for IRT and STST were 0.024 and 0.021, respectively; for conventional equipercentile and linear methods, these values were 0.036 and 0.048, respectively. STST had smaller bias than any of the other three methods. Nonparallel power subtests were equated with greater error than were parallel power subtests.

Linear equating methods worked best for equating nonparallel speeded subtests (RMSE equal to 0.006), with STST methods performing the most poorly (RMSE equal to 0.015). There were small differences across the methods in bias. Parallel and nonparallel speeded tests were equated equally well. Because nonparallel speeded tests differed in length but not in difficulty, this may suggest that varying difficulty has more of an effect on equating than does varying test length. This issue will be discussed in more detail later.

Discussion

Conventional equating methods outperformed the more complex IRT and STST methods when parallel subtests were equated. In fact, the simplest (linear) method worked much better than any of the other methods when speeded subtests were equated. Exactly the opposite was true when nonparallel power subtests were equated, however. That is, IRT and STST clearly worked better than the conventional methods; STST yielded a smaller bias than all other methods. In general, parallel subtests were equated with less error than were nonparallel subtests. As one exception, however, nonparallel speeded tests were equated with the same amount of error as the parallel speeded subtests, suggesting that variation in test length alone was not a significant violation of test parallelism.

It appears, then, that the conventional equating methods function well when parallel tests are equated but work less well than IRT and STST methods for equating nonparallel tests.

Data Collection Designs

Results

Table 21 presents the true-score error indices (for each data collection design) for equating parallel subtests. Differences across data collection designs were small. In general, the single-group design resulted in smaller RMSEs than did the equivalent-groups and anchor-test designs; this was especially true when speeded subtests were equated (the pooled RMSEs were 0.008, 0.010, and 0.009, respectively, for the power subtests and 0.006, 0.011, and 0.013, respectively, for the speeded subtests). There were essentially no differences in the errors yielded by the equivalent-groups and anchor-test designs. In general, bias was small throughout; the single exception to this occurred for equipercentile anchor-test equating, which resulted in a positive bias for the speeded subtests.

Table 21
True-Score Error Indices for Equating Parallel Subtests Using Different Data Collection Designs

	Data collection designs						
	Single group		Equivalent groups		Anchor test		
Equating method	RMSE	Bias	RMSE	Bias	RMSE	Bias	N of equatings
Power subtests							
Linear	0.006	0.001	0.007	0.001	0.007	-0.001	36
Equipercentile	0.007	0.001	0.008	0.001	0.009	-0.001	36
IRT	0.008	0.001	0.012	0.001	0.009	-0.002	36
STST	0.010	0.001	0.011	0.002	0.012	-0.001	36
Pooled	0.008	0.001	0.010	0.001	0.009	-0.001	144
Speeded subtests							
Linear	0.002	0.000	0.007	-0.003	0.003	-0.001	12
Equipercentile	0.006	0.000	0.009	-0.003	0.013	0.006	12
STST	0.008	-0.001	0.016	-0.002	0.019	-0.003	12
Pooled	0.006	-0.001	0.011	-0.003	0.013	0.001	36

Table 22 presents the true-score error indices computed when nonparallel subtests were equated. Error indices for the nonparallel subtests were generally larger than those observed when parallel subtests were equated. When nonparallel power subtests were equated, only small mean differences among the data collection designs were evident, except for IRT equating; here the single-group design was best and the equivalent-groups design was the worst (RMSEs of 0.012 and 0.032, respectively). Bias was large (0.013) for IRT using the anchor-test design. Overall, there was a slight advantage for the single-group design (mean pooled RMSE of 0.032 vs. 0.034-0.036).

Table 22
True-Score Error Indices for Equating Nonparallel Subtests Using Different Data Collection Designs

	Data collection design						
	Single group		Equivalent groups		Anchor test		
Equating method	RMSE	Bias	RMSE	Bias	RMSE	Bias	N of equatings
Power subtests							
Linear	0.047	0.007	0.049	0.007	0.048	0.005	45
Equipercentile	0.035	0.010	0.036	0.009	0.036	0.007	45
IRT	0.012	0.003	0.032	-0.002	0.024	0.013	45
STST	0.020	0.000	0.020	0.000	0.022	-0.003	45
Pooled	0.032	0.005	0.036	0.003	0.034	0.006	180
Speeded subtests							
Linear	0.004	0.001	0.008	-0.002	0.005	0.000	15
Equipercentile	0.006	0.001	0.011	-0.003	0.018	0.009	15
STST	0.010	-0.001	0.017	-0.002	0.016	-0.002	15
Pooled	0.007	0.000	0.013	-0.002	0.014	0.002	45

For the speeded subtests, however, differences among designs were more marked: The single-group design was consistently the best design and linear equating was the best method. Bias was largest (0.009) for equipercentile equating using anchor tests. No consistent differences were observed between the equivalent-groups and anchor-test designs. As was discussed previously, nonparallel speeded subtests were equated with approximately the same degree of error as were the parallel speeded subtests, again suggesting that test length was not a major factor contributing to the error in nonparallel-speeded-test pairings.

Discussion

In general, the single-group data collection design was clearly best for equating nonparallel power subtests using IRT and for equating speeded subtests by any of the three equating methods. These were the only cases in which a data collection design was clearly superior for subtest equating.

The clear superiority of IRT single-group equating over any other type of IRT equating is most probably due to the particular implementation of that data collection design with item response theory. For all other equating methods, the single-group and equivalent-groups designs differ only in that the latter design uses two separate samples of examinees (instead of just one) to obtain the equating transformation. Differences between these two designs, then, arise from the additional sampling error that is involved in the equivalent-groups design.

For IRT single-group equating, however, all items on both subtests are simultaneously calibrated; item calibration for IRT equivalent-groups equating is performed separately for each subtest and each examinee sample. It has been demonstrated (e.g., Vale et al., 1981) that increased item set size yields better parameter estimates for all the items. Better parameter estimates, in turn, yield a more accurate IRT equating transformation. Hence, the single-group design as implemented with IRT has two advantages over the equivalent-groups design: (a) smaller sampling error, and (b) better item parameter estimates.

Item response theory had a lower RMSE but larger bias with an anchor test than it did when equivalent groups were used. When speeded subtests were equated using equipercentile procedures, both error indices were higher for the anchor-test design than for the equivalent-groups design. Linear anchor-test equating was clearly superior to linear equivalent-groups equating only for speeded subtests. Anchor-test equating using STST was typically slightly worse than equivalent-groups STST equating.

These results can perhaps be best explained by recalling the definition of anchor-test equating used in this study. For the conventional and STST methods, scores on each of the two tests were first equated to a separate anchor test. Scores that were equated to the same anchor-test score were considered to be equated to each other. In order to equate two sets of scores, then, two separate and independent equatings were performed. It is likely that equating error was compounded; this could account for the fact that anchor-test equating was usually worse than equivalent-groups equating for equipercentile and STST methods; results for the linear procedure were equivocal. Because the examinee groups were defined to be equivalent in ability, the anchor-test design provided no tangible benefit for these methods.

For item response theory, however, the anchor-test design was implemented in a slightly different way. Items on both tests and the anchor test were simultaneously calibrated, putting all item parameter and ability estimates on the same scale. This probably yielded better item parameter estimates and therefore better equating (at least in terms of RMSE) than did the equivalent-groups design and its assumption of exactly equivalent true-ability distributions.

Sample Sizes

Results

Table 23 presents the true-score error indices for equating parallel subtests using various sample sizes. This table reveals that there was a drop in the pooled RMSE as sample size increased from 1,000 to 2,400 (from 0.011 to 0.007 for both power and speeded subtests). For the power subtests, RMSEs computed from the selected sample generally fell between the values for the unselected samples; for the speeded subtests, RMSEs were highest for the selected samples. Bias was small throughout. These patterns were consistent across all equating methods.

Table 23
True-Score Error Indices for Equating Parallel Subtests Using Various Sample Sizes

	Sample size								
	1000 (unselected)			1600 (selected)			2400 (unselected)		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*
Power subtests									
Linear	0.009	-0.001	36	0.006	0.001	36	0.006	0.000	36
Equipercentile	0.010	-0.001	36	0.007	0.001	36	0.006	0.000	36
IRT	0.011	-0.001	36	0.010	0.000	36	0.008	0.000	36
STST	0.013	-0.001	36	0.011	0.001	36	0.009	0.001	36
Pooled	0.011	-0.001	144	0.009	0.001	144	0.007	0.001	144
Speeded subtests									
Linear	0.006	-0.001	12	0.003	-0.001	12	0.003	-0.001	12
Equipercentile	0.009	-0.001	12	0.013	0.003	12	0.007	0.001	12
STST	0.016	-0.003	12	0.018	-0.002	12	0.009	-0.002	12
Pooled	0.011	-0.002	36	0.013	0.000	36	0.007	-0.001	36

*Number of equating tables included in the pooled error indices.

Table 24 presents similar indices for the nonparallel-test equatings. When nonparallel power subtests were equated, there was little decrease in pooled RMSE (from 0.031 to 0.029) as sample size increased from 1,000 to 2,400, but a large increase (to 0.041) when a selected sample was used; bias increased from 0.001 to 0.013 for the selected sample. Most of this increase can be attributed to the conventional equating methods. IRT equating was only slightly affected by this manipulation; STST was robust against the use of selected examinee samples.

Table 24
True-Score Error Indices for Equating Nonparallel Subtests Using Various Sample Sizes

Equating method	Sample size 1000 (unselected) RMSE	 Bias	 N*	 1600 (selected) RMSE	 Bias	 N*	 2400 (unselected) RMSE	 Bias	 N*
Power subtests									
Linear	0.043	-0.001	45	0.058	0.022	45	0.042	-0.001	45
Equipercentile	0.029	0.001	45	0.047	0.022	45	0.028	0.002	45
IRT	0.025	0.004	45	0.026	0.006	45	0.022	0.004	45
STST	0.021	-0.002	45	0.020	0.000	45	0.020	-0.002	45
Pooled	0.031	0.001	180	0.041	0.013	180	0.029	0.001	180
Speeded subtests									
Linear	0.007	-0.002	15	0.007	0.001	15	0.004	-0.001	15
Equipercentile	0.012	0.001	15	0.015	0.004	15	0.009	0.002	15
STST	0.017	-0.003	15	0.018	-0.001	15	0.008	-0.001	15
Pooled	0.013	-0.001	45	0.014	0.001	45	0.008	0.000	45

*Number of equating tables included in the pooled error indices.

For the nonparallel speeded subtests there was a decrease in the pooled RMSE (from 0.013 to 0.008) as sample size increased, and only a slight increase (to 0.014) when a selected sample was used. These patterns were consistent across equating methods. In general, the discrepancy in the error indices between the selected and unselected samples was much larger for the power subtests than it was for the speeded subtests.

Discussion

Increasing the sample size from 1,000 to 2,400 examinees had only a small effect on equating accuracy; for nonparallel power subtests, the effect was negligible. The use of selected examinee samples did not greatly affect equating accuracy for pairs of parallel subtests and for nonparallel speeded subtests (which, for all practical purposes, have been behaving like parallel subtests). These patterns were consistent across all equating methods.

When nonparallel power subtests were conventionally equated using selected examinee samples, however, both the RMSE and bias increased substantially. IRT equating was only slightly affected by the use of selected samples; STST equating was not affected at all.

Test Lengths and Difficulties

Results

Table 25 presents the true-score error indices for parallel-test equatings when test difficulty and length were varied. In general, both subtest length and difficulty had a minor effect on the accuracy of test equating. For the easy power subtests, pooled RMSE decreased from 0.011 to 0.008 as test length increased from 15 to 30 items; similarly, mean bias decreased from 0.004 to 0.001. For the difficult power subtests, the decrease in error was even smaller. Bias was consistently small and positive when easy power subtests were equated and was small and negative when difficult power tests were equated. For the speeded subtests, the effect of test length on equating accuracy was negligible.

Table 25
True-Score Error Indices for Equating Parallel Subtests Using Various Levels of Test Difficulty and Length

	Subtest difficulty					
	Easy			Difficult		
Test length	RMSE	Bias	N*	RMSE	Bias	N*
Power subtests						
Short	0.011	0.004	108	0.009	-0.003	108
Long	0.008	0.001	108	0.008	-0.001	108
Speeded subtests						
Short	0.011	-0.002	54	-	-	-
Long	0.010	0.000	54	-	-	-

Note. Difficulty was not explicitly varied for the speeded subtests. Hence, the error indices for all speeded-test forms were pooled for this table.
*Number of equating tables included in pooled error estimates.

The error indices for both power and speeded subtests were pooled and are presented, separately by equating method, in Table 26. In general, the error patterns were consistent for all equating methods. That is, RMSE and bias decreased slightly when subtest length was increased. IRT showed the largest decrease in RMSE, from 0.013 to 0.007 for the easy subtests. RMSE indices were not dramatically affected by subtest difficulty. The effect on bias was consistent though small; there was a slight positive bias when easy subtests were used and a slight negative bias when difficult subtests were used. All methods were essentially unbiased at the longer test lengths.

Table 26
True-Score Error Indices for Equating Parallel Subtests Using Different Equating Methods and Various Levels of Test Length and Difficulty

	Subtest difficulty					
	Easy			Difficult		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*
Short subtests						
Linear	0.008	0.003	36	0.007	-0.003	36
Equipercentile	0.010	0.003	36	0.008	-0.002	36
IRT	0.013	0.004	27	0.010	-0.003	27
STST	0.012	0.003	36	0.013	-0.003	36
Pooled	0.011	0.003	135	0.010	-0.003	135
Long subtests						
Linear	0.005	0.001	36	0.004	-0.001	36
Equipercentile	0.008	0.001	36	0.008	-0.001	36
IRT	0.007	-0.001	27	0.008	-0.001	27
STST	0.011	0.001	36	0.012	-0.001	36
Pooled	0.008	0.000	135	0.009	-0.001	135

Table 27 presents the true-score error indices resulting when nonparallel subtests were equated using differing levels of test length and/or difficulty. The pooled error indices are presented for subtest pairings where the tests that were equated were of (a) different difficulty but equal length, (b) different length but equal difficulty, and (c) both different length and difficulty. These columns correspond to test pairings five through nine, respectively.

The test-length effect that was evident (though slight) for the parallel subtests was more marked for nonparallel power subtests. That is, the RMSE (pooled over all equating methods) decreased from 0.043 to 0.036 as test length increased; similarly, pooled bias decreased from 0.007 to 0.005. This same pattern was evident for all the equating methods and was largest for the equipercentile equating procedure and smallest for linear. There was essentially no test-length effect for the speeded subtests.

Varying difficulty level across the subtests being equated (as was done for the first four columns of Table 27) resulted in rather large true-score error indices for the conventional methods (RMSEs between 0.033 and 0.060) and somewhat smaller indices for IRT and STST (RMSEs between 0.018 and 0.032). In general, errors for this case of vertical equating were much smaller for IRT and STST than they were for the conventional methods.

Table 27

True-Score Error Indices for Equating Nonparallel Subtests Using Different Equating Methods and Various Levels of Test Length and Difficulty

	Different difficulty				Different length				Different length and difficulty		N of equatings per cell
	Short		Long		Easy		Difficult				
Equating method	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	
Power subtests											
Linear	0.060	0.010	0.056	0.007	0.016	0.004	0.016	0.001	0.066	0.011	27
Equipercentile	0.045	0.016	0.033	0.010	0.017	0.003	0.017	-0.001	0.052	0.013	27
IRT	0.032	0.005	0.025	0.005	0.017	0.007	0.013	-0.001	0.030	0.008	27
STST	0.027	-0.003	0.018	-0.001	0.014	0.003	0.015	0.000	0.025	-0.005	27
Pooled	0.043	0.007	0.036	0.005	0.016	0.004	0.015	0.000	0.046	0.007	108
Speeded subtests											
Linear	0.005	-0.002	0.003	0.000	0.008	0.000	-	-	0.007	-0.001	9
Equipercentile	0.011	0.000	0.011	-0.001	0.014	0.004	-	-	0.012	0.004	9
STST	0.017	-0.004	0.014	-0.001	0.015	-0.001	-	-	0.014	-0.002	9
Pooled	0.012	-0.002	0.011	-0.001	0.012	0.001	-	-	0.011	0.000	27

Note. Difficulty was not explicitly varied for the speeded subtests. Hence, the error indices for all speeded-subtest forms were pooled for the easy-difficult contrast (columns 5-8). The number of equatings for that group of cells is therefore twice as large as the number indicated in the last column of the table.

Equating accuracy was not differentially affected by the level of difficulty of the individual subtests. That is, pooled error estimates were essentially identical for the two different difficulty levels investigated in this study. This was true for the individual equating methods and for all methods overall; only IRT equating yielded a smaller RMSE for the difficult subtests (0.013 vs. 0.017).

Varying both the length and difficulty across subtests resulted in the highest error indices observed for the conventional equating methods (RMSEs of 0.066 for linear and 0.052 for equipercentile). It resulted in indices for IRT and STST methods of about the same magnitude as were observed when tests that differed only in difficulty were equated.

Discussion

When parallel subtests were equated, there were only minor effects on equating accuracy that could be attributed to subtest length and difficulty. That is, there was a slight decrease in the RMSE and bias when test length was doubled. There were essentially no differences between equating easy-to-easy and difficult-to-difficult subtests. Equated scores derived from easy subtests yielded a small positive bias at the shorter subtest lengths; equating transformations derived from short difficult subtests yielded a small negative bias. This effect was consistent for all the equating methods.

Similar test-length effects were evident when nonparallel subtests were equated. Subtests that varied in difficulty level were equated with a much greater degree of error than were subtests that differed only in length. This was true for all the equating methods but particularly so for the conventional methods. This finding suggests caution when tests of different difficulty are to be (vertically) equated, and that conventional equating methods should not be used in this situation.

Ability Levels

Results

Equating using different ability levels (one higher-ability sample and one lower-ability sample), as well as equivalent ability levels, was performed for Subtest PC using the equivalent-groups and anchor-test designs. Table 28 presents the results for the different equating methods and data collection designs when (a) examinee ability distributions were equivalent and (b) examinees differed in mean ability.

As was observed earlier, there were essentially no differences between the equivalent-groups and anchor-test designs when the groups

Table 28
True-Score Error Indices for Equating Across Ability Levels on Subtest PC

	Equivalent ability levels						Different ability levels					
	Equivalent groups			Anchor test			Equivalent groups			Anchor test		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*
Parallel subtests												
Linear	0.006	0.001	12	0.007	-0.003	12	0.027	-0.024	8	0.010	-0.002	8
Equipercentile	0.008	0.001	12	0.008	-0.002	12	0.027	-0.024	8	0.011	-0.002	8
IRT	0.012	0.002	12	0.008	-0.002	12	0.026	-0.021	8	0.013	-0.002	8
STST	0.009	0.001	12	0.011	-0.004	12	0.029	-0.024	8	0.012	-0.002	8
Pooled	0.009	0.001	48	0.009	-0.003	48	0.027	-0.023	32	0.011	-0.002	32
Nonparallel subtests												
Linear	0.031	0.005	15	0.030	0.003	15	0.034	-0.023	10	0.026	-0.003	10
Equipercentile	0.024	0.007	15	0.023	0.005	15	0.028	-0.021	10	0.015	-0.001	10
IRT	0.014	0.001	15	0.024	0.012	15	0.030	-0.023	10	0.017	0.008	10
STST	0.011	0.000	15	0.014	-0.003	15	0.031	-0.026	10	0.018	-0.006	10
Pooled	0.022	0.003	60	0.024	0.004	60	0.031	-0.023	40	0.019	0.000	40

*Number of equatings included in the pooled error estimates.

were of equivalent ability. There was a small but consistent negative bias when an anchor test was used to equate parallel subtests. However, when the groups differed in ability level, the anchor-test design produced much better equating in terms of smaller RMSEs and smaller bias indices. Anchor-test equating using samples of different ability had only slightly higher RMSE than when it was applied to samples of equivalent ability. Equivalent-groups equating had higher RMSE and bias when the ability levels of the groups differed; this should be expected since a major assumption of the design was violated. There were essentially no differences in RMSE across the equating methods when parallel tests were equated using different ability levels.

Similar findings were evident when nonparallel tests were equated. Few differences were observed between the two data collection designs when the ability levels were equivalent. As before, the anchor-test design produced much better equating when the groups differed in ability levels. Anchor-test equating actually yielded slightly lower pooled RMSEs when ability levels were different than when they were the same (0.019 vs. 0.024). Equivalent-groups equating again had higher errors when the ability levels of the groups differed.

Discussion

When nonequivalent examinee samples were used to equate two subtests, the anchor-test data collection design consistently yielded lower indices of equating error. This was true for all equating methods and for both parallel and nonparallel subtests. In fact, anchor-test equating was typically as accurate using nonequivalent examinee samples as it was using equivalent samples.

Equating Test Composites

Power and AFQT composites were equated (a) directly, (b) indirectly through the subtests, and (c) by forming composites of equated subtests. Equating transformations derived using these various methods were evaluated separately; results are presented below.

Equating Methods

Results

Strong true-score theory. The STST procedures used in this simulation project were developed from Lord's published descriptions of his methods (Lord, 1965, 1969). However, severe computer representation and overflow problems were encountered when STST was applied to composites that contained as many as 90 items; these problems could not be solved within the time frame allowed for this project. Lord's original computer programs (Stocking, Wingersky, Lees, Lennon, & Lord, 1973; Wingersky, Lees, Lennon, & Lord, 1969) were not used for this project because they were not readily adaptable to these simulations; it is possible that these implementations of strong true-score theory contain refinements to the procedures that are able to overcome some of these numerical difficulties. However, it should be noted that Stocking et al. (1973) limit the number of test items to 50 in their program.

Because of these numerical difficulties, STST was applied only to composites that contained 45 items; this included pairings 1, 3, and 5.

Equating composite scores directly. Table 29 presents the true-score error indices for equating composite scores directly. As was observed when individual subtests were equated, parallel composites were equated with substantially less error than were the nonparallel composites. For the parallel and nonparallel power composites, the pooled RMSEs were 0.008 and 0.031, respectively; for the AFQT composites, these figures were 0.007 and 0.025. The same pattern was observed for the pooled bias indices.

Table 29
True-Score Error Indices for Equating Composite Scores Directly

	Type of composite								
	Power			AFQT			Power to AFQT		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*
Parallel composites									
Linear	0.005	0.001	36	0.004	0.001	24	0.035	0.007	36
Equipercentile	0.009	0.001	36	0.008	0.001	24	0.037	0.009	36
STST	0.010	0.001	12	-	-	-	-	-	-
Pooled	0.008	0.001	84	0.007	0.001	48	0.036	0.008	72
Nonparallel composites									
Linear	0.037	0.007	45	0.029	0.004	30	0.052	0.012	45
Equipercentile	0.026	0.008	45	0.021	0.004	30	0.046	0.014	45
STST	0.020	0.003	6	-	-	-	-	-	-
Pooled	0.031	0.007	96	0.025	0.004	60	0.049	0.013	90

Note. STST was applied only to short (45-item) composites.
*Number of equating tables included in the pooled error estimates.

For both the parallel- and nonparallel-composite pairings, there were consistent differences between linear and equipercentile methods. For the parallel power and AFQT composites, smaller RMSEs were observed for the linear equating method (0.005 vs. 0.009 for the power composites and 0.004 vs. 0.008 for the AFQT composites), whereas equipercentile equating was better for the nonparallel composites (0.026 vs. 0.037 for the power composites and 0.021 vs. 0.029 for the AFQT composites). RMSE for STST was equal to 0.010 for the parallel composites (larger than either of the conventional methods) and was equal to 0.020 for the nonparallel composites. This latter value for STST equating of nonparallel composites was based only on pairing 5, involving the shorter test lengths. For this single pairing, RMSE was equal to 0.046 and 0.033 for the linear and equipercentile methods, respectively.

Bias was negligible (0.001) for the parallel power and AFQT composites and somewhat larger (0.003-0.008) for the nonparallel composites. Again, STST resulted in a smaller error than did the conventional methods for equating nonparallel composites. For pairing 5 only, bias was equal to 0.011 and 0.014 for the linear and equipercentile methods, respectively; these values compare with 0.003 for STST.

The pooled error indices for the power composite were larger than those for the AFQT composite for the nonparallel composite forms; no such effect was observed for the parallel composite forms.

Equating unlike composites (power to AFQT) resulted in considerable RMSE (up to 0.052 for the nonparallel, linear case) and a positive bias. Again, linear methods worked slightly better than equipercentile methods for parallel forms (RMSE of 0.035 vs. 0.037) and were somewhat worse for nonparallel forms (RMSE of 0.052 vs. 0.046). The errors observed when unlike composites were equated were much larger than those observed for either of the other two composite types.

Forming composites of equated subtests. Table 30 presents the true-score error indices computed from composites of equated subtests. There were few differences observed across the four equating methods when parallel composites were equated. That is, the levels of RMSE for all the equating methods were essentially the same (0.004-0.006), and all methods were unbiased.

When nonparallel power composites were equated, however, there was a moderate degree of bias for the conventional (0.006-0.008) and IRT methods (0.004); STST was essentially unbiased. Conventional equating methods yielded larger RMSEs (0.031-0.041) than did IRT (0.019) or STST (0.014) methods. The same pattern of errors was observed for the nonparallel AFQT composites. Again, AFQT composites were equated with less error than were the power composites.

Table 30
True-Score Error Indices for Forming Composites of Equated Subtests

	Type of composite					
	Power			AFQT		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*
Parallel composites						
Linear	0.004	0.000	36	0.004	0.000	36
Equipercentile	0.005	0.000	36	0.004	0.000	36
IRT	0.005	0.000	36	-	-	
STST	0.006	0.001	36	0.006	0.000	36
Pooled	0.005	0.000	144	0.005	0.000	108
Nonparallel composites						
Linear	0.041	0.006	45	0.032	0.005	45
Equipercentile	0.031	0.008	45	0.025	0.007	45
IRT	0.019	0.004	45	-	-	
STST	0.014	-0.001	45	0.012	-0.001	45
Pooled	0.028	0.005	180	0.024	0.004	135

*Number of equating tables included in the pooled error estimates.

For the parallel case, forming composites from previously equated subtests resulted in slightly smaller amounts of equating error than did equating composite scores directly. For the nonparallel case, however, equating composite scores directly using conventional methods resulted in slightly lower RMSEs than did forming composites from conventionally equated subtests (RMSEs of 0.026-0.037 vs .0.031-0.041). When STST was used, directly equated subtests had larger errors than did the composites formed from previously equated subtests.

Equating composite scores indirectly through the subtests. Indirect methods of equating composite scores were developed for the case in which only partial data are available. Table 31 shows that this method using partial data performed as well as the other methods of equating composites (see Tables 29 and 30). However, it should be noted that only the larger sample sizes were used when composites were indirectly equated; Tables 29 and 30 include the results from all sample sizes. There were no effects on equating accuracy that could be attributed to data collection design. The use of selected examinee samples resulted in larger RMSE and bias for the nonparallel composites; there was no such effect evident for the parallel composites. Again, AFQT composites were equated with less error than were the power composites.

Table 31

True-Score Error Indices for Equating Composites Indirectly Through the Subtests

	Type of composite					
	Power			AFQT		
Data collection design	RMSE	Bias	N*	RMSE	Bias	N*
Parallel composites						
Single group (N=2,400)	0.004	0.000	4	0.004	0.000	4
Equivalent groups (N=1,600)	0.003	0.000	4	0.004	-0.002	4
Equivalent groups (N̄=2,400)	0.003	0.000	4	0.003	-0.001	4
Nonparallel composites						
Single group (N=2,400)	0.030	-0.002	5	0.024	0.000	5
Equivalent groups (N=1,600)	0.045	0.017	5	0.038	0.015	5
Equivalent groups (N̄=2,400)	0.030	-0.001	5	0.023	0.000	5

*Number of equating tables included in the pooled error estimates.

Discussion

In general, only small differences were observed among the equating methods used for test composites. When composite scores were directly equated, linear methods worked better for the parallel composites, and STST and equipercentile methods worked better for the nonparallel composites. When composites were formed from equated subtests, differences among equating methods were observed only for the nonparallel composites: Conventional methods resulted in higher RMSEs and greater bias than did IRT and STST methods; STST was unbiased. Comparison of Tables 29 and 30 reveals that there was slightly less error involved when composites were formed from equated subtests than when they were directly equated. The use of an indirect equating procedure, with only a subset of the examinee response data, did not adversely affect equating accuracy.

Data Collection Designs

Results

Table 32 shows that there were only minor differences among the data collection designs when parallel composites were equated. The pooled RMSEs for each of the data collection designs were essentially identical within each type of composite. The single exception to this ocurred for the direct power composites, where anchor-test equating was slightly worse than was equating using any of the other designs. This difference can be attributed to the fact that the RMSE was higher for equipercentile anchor-test equating than it was for any other method using the anchor-test design. In all other cases, equating methods performed consistently across data collection designs.

Table 32
True-Score Error Indices for Equating Parallel Composites Using Different Data Collection Designs

	Data collection designs								
	Single group			Equivalent groups			Anchor test		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*
Direct power									
Linear	0.004	0.001	12	0.005	0.001	12	0.006	0.000	12
Equipercentile	0.007	0.002	12	0.008	0.001	12	0.011	0.001	12
STST	0.009	0.002	6	0.011	0.001	-	-	-	-
Pooled	0.007	0.001	30	0.008	0.001	30	0.009	0.001	24
Direct AFQT									
Linear	0.004	0.001	12	0.004	0.000	12	-	-	
Equipercentile	0.008	0.001	12	0.008	0.001	12	-	-	
Pooled	0.006	0.001	24	0.007	0.000	24	-	-	
Direct power to AFQT									
Linear	0.035	0.008	12	0.035	0.007	12	0.035	0.007	12
Equipercentile	0.036	0.009	12	0.036	0.009	12	0.037	0.009	12
Pooled	0.036	0.008	24	0.036	0.008	24	0.036	0.008	24
Equated power subtests									
Linear	0.004	0.001	12	0.004	0.001	12	0.005	-0.001	12
Equipercentile	0.004	0.001	12	0.005	0.001	12	0.005	-0.001	12
IRT	0.005	0.001	12	0.006	0.000	12	0.006	-0.002	12
STST	0.006	0.001	12	0.006	0.002	12	0.007	-0.001	12
Pooled	0.005	0.001	48	0.005	0.001	48	0.006	-0.001	48
Equated AFQT subtests									
Linear	0.003	0.001	12	0.004	0.000	12	0.004	-0.001	12
Equipercentile	0.004	0.000	12	0.004	0.000	12	0.005	0.001	12
STST	0.005	0.001	12	0.006	0.001	12	0.007	-0.002	12
Pooled	0.004	0.001	36	0.005	0.000	36	0.005	-0.001	36

*Number of equating tables included in the pooled error estimates.

As was observed earlier, there was no difference in equating accuracy between the direct power and AFQT composites for the parallel forms. There was a much greater amount of error involved when a power composite was equated directly to a composite of AFQT subtests; this was the only instance in which a consistent positive bias was observed. Forming composites from equated subtests yielded slightly less error than did equating composite scores directly; this was equally true for all the data collection designs.

Table 33 presents the error indices computed when nonparallel composites were equated using different data collection designs. None of the data collection designs was consistently best for equating nonparallel composites. AFQT composites were equated with less error

Table 33
True-Score Error Indices for Equating Nonparallel Composites Using Different Data Collection Designs

	Data collection designs								
	Single group			Equivalent groups			Anchor test		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*
Direct power									
Linear	0.036	0.007	15	0.038	0.006	15	0.037	0.006	15
Equipercentile	0.025	0.008	15	0.026	0.008	15	0.027	0.008	15
STST	0.021	0.002	3	0.020	0.003	3	-	-	-
Pooled	0.030	0.007	33	0.032	0.007	33	0.032	0.007	30
Direct AFQT									
Linear	0.028	0.004	15	0.029	0.003	15	-	-	
Equipercentile	0.020	0.005	15	0.023	0.004	15	-	-	
Pooled	0.025	0.005	30	0.026	0.004	30	-	-	
Direct power to AFQT									
Linear	0.052	0.012	15	0.053	0.011	15	0.052	0.011	15
Equipercentile	0.046	0.014	15	0.047	0.014	15	0.047	0.014	15
Pooled	0.049	0.013	30	0.050	0.012	30	0.050	0.013	30
Equated power subtests									
Linear	0.040	0.007	15	0.041	0.007	15	0.041	0.005	15
Equipercentile	0.030	0.010	15	0.032	0.009	15	0.031	0.007	15
IRT	0.009	0.003	15	0.023	-0.003	15	0.022	0.003	15
STST	0.014	0.000	15	0.015	0.000	15	0.014	-0.003	15
Pooled	0.026	0.005	60	0.029	0.003	60	0.029	0.006	60
Equated AFQT subtests									
Linear	0.032	0.006	15	0.033	0.005	15	0.032	0.004	15
Equipercentile	0.024	0.008	15	0.025	0.006	15	0.024	0.007	15
STST	0.011	0.000	15	0.012	-0.001	15	0.012	-0.003	15
Pooled	0.024	0.004	45	0.025	0.003	45	0.024	0.003	45

*Number of equating tables included in the pooled error indices.

than were power composites. Again, the direct equating of a power to an AFQT composite resulted in a consistent positive bias. Moderate levels of bias were observed throughout for the conventional and IRT equating methods; STST equating was essentially unbiased. All equating methods performed consistently across data collection designs with the single exception that forming composites of IRT-equated subtests using the single-group design yielded a lower RMSE (0.009) than did any other design in conjuntion with IRT (0.022-0.023).

Discussion

None of the data collection designs proved to be consistently best or, for that matter, consistently worst for equating composites of any type. With few exceptions, the equating methods performed consistently across the different data collection designs; there was no distinct method-by-design interaction. STST was essentially unbiased for those conditions where it was applied.

Sample Sizes

Results

Table 34 presents the true-score error indices computed when parallel composites were equated using various sample sizes. When power and AFQT composites were directly equated, there was a minor effect on equating accuracy that could be attributed to increasing the examinee sample size from 1,000 to 2,400. For the power composites, pooled RMSE decreased from 0.009 to 0.005; for the AFQT composites, these figures were 0.006 and 0.003, respectively. When a power composite was directly equated to an AFQT composite, there was no advantage to using the larger sample size; pooled RMSE and bias were equal to 0.033 and 0.004, respectively, for both examinee groups.

Direct linear equating of composites resulted in slightly smaller error indices than did equipercentile and STST composite equating, particularly when the smaller sample size was used. For the power composites, the RMSEs were 0.006, 0.010, and 0.013, respectively, for the smaller sample size. For the AFQT composite, these figures were 0.005 and 0.007 for the linear and equipercentile methods, respectively.

The use of a selected examinee sample did not affect the accuracy of equating power composites directly. When AFQT composites were directly equated, however, error increased for equipercentile equating: RMSE for the selected sample was 0.012, compared to 0.007 and 0.004 for the smaller and larger samples, respectively; bias also increased slightly. There was no corresponding effect when the AFQT composites were linearly equated.

Table 34
True-Score Error Indices for Equating Parallel Composites Using Various Sample Sizes

	Sample size								
	1000 (unselected)			1600 (selected)			2400 (unselected)		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*
Direct power									
Linear	0.006	0.000	12	0.005	0.002	12	0.004	0.000	12
Equipercentile	0.010	0.001	12	0.010	0.003	12	0.005	0.001	12
STST	0.013	0.001	4	0.009	0.002	4	0.007	0.001	4
Pooled	0.009	0.001	28	0.008	0.002	28	0.005	0.001	28
Direct AFQT									
Linear	0.005	-0.001	8	0.004	0.002	8	0.003	0.001	8
Equipercentile	0.007	-0.001	8	0.012	0.003	8	0.004	0.001	8
Pooled	0.006	-0.001	16	0.009	0.002	16	0.003	0.001	16
Direct power to AFQT									
Linear	0.033	0.003	12	0.038	0.016	12	0.033	0.003	12
Equipercentile	0.033	0.004	12	0.043	0.019	12	0.032	0.004	12
Pooled	0.033	0.004	24	0.041	0.017	24	0.033	0.004	24
Equated power subtests									
Linear	0.006	-0.001	12	0.003	0.001	12	0.003	0.000	12
Equipercentile	0.006	-0.001	12	0.004	0.001	12	0.004	0.000	12
IRT	0.007	-0.001	12	0.005	0.000	12	0.004	0.000	12
STST	0.008	-0.001	12	0.006	0.001	12	0.005	0.001	12
Pooled	0.007	-0.001	48	0.005	0.001	12	0.004	0.000	48
Equated AFQT subtests									
Linear	0.005	-0.001	12	0.002	0.000	12	0.003	0.000	12
Equipercentile	0.005	-0.001	12	0.004	0.001	12	0.003	0.001	12
STST	0.007	-0.001	12	0.006	0.001	12	0.004	0.000	12
Pooled	0.006	-0.001	36	0.005	0.001	36	0.003	0.000	36

*Number of equating tables included in the pooled error estimates.

Using a selected examinee sample to equate a power composite to an AFQT composite caused a substantial increase in equating error. This was the only situation in which bias was large and positive; in all other cases, bias was essentially zero. Equating error increased for both equating methods, but especially so for equipercentile, where RMSE increased from 0.032-0.033 to 0.043.

When parallel composites were formed from equated subtests, there was a slight sample-size effect; this was true for all of the equating methods. Pooled RMSE decreased from 0.007 to 0.004 for the power composites and from 0.006 to 0.003 for the AFQT composites. The use of a selected examinee sample did not affect equating accuracy.

Table 35 presents the error indices computed when nonparallel composites were equated using various sample sizes. Increasing sample size from 1,000 to 2,400 had little effect on equating accuracy; this was true for all equating methods and for all types of composite equating.

Table 35
True-Score Error Indices for Equating Nonparallel Composites Using Various Sample Sizes

	Sample size								
	1000 (unselected)			1600 (selected)			2400 (unselected)		
Equating method	RMSE	Bias	N*	RMSE	Bias	N*	RMSE	Bias	N*
Direct power									
Linear	0.031	0.000	15	0.047	0.020	15	0.030	0.000	15
Equipercentile	0.019	0.003	15	0.037	0.018	15	0.018	0.002	15
STST	0.021	0.003	2	0.020	0.004	2	0.020	0.002	2
Pooled	0.025	0.002	32	0.041	0.018	32	0.024	0.001	32
Direct AFQT									
Linear	0.024	-0.002	10	0.037	0.015	10	0.023	-0.001	10
Equipercentile	0.016	-0.001	10	0.030	0.013	10	0.015	0.001	10
Pooled	0.021	-0.002	20	0.034	0.014	20	0.020	0.000	20
Direct power to AFQT									
Linear	0.045	0.002	15	0.064	0.031	15	0.045	0.002	15
Equipercentile	0.038	0.006	15	0.060	0.030	15	0.037	0.005	15
Pooled	0.042	0.004	30	0.062	0.030	30	0.041	0.004	30
Equated power subtests									
Linear	0.035	-0.001	15	0.051	0.022	15	0.034	-0.001	15
Equipercentile	0.024	0.001	15	0.042	0.022	15	0.023	0.002	15
IRT	0.020	0.004	15	0.019	0.006	15	0.018	0.003	15
STST	0.014	-0.002	15	0.014	0.000	15	0.015	-0.002	15
Pooled	0.024	0.001	60	0.035	0.013	60	0.023	0.001	60
Equated AFQT subtests									
Linear	0.027	-0.001	15	0.041	0.017	15	0.027	-0.001	15
Equipercentile	0.019	0.001	15	0.034	0.018	15	0.018	0.002	15
STST	0.011	-0.002	15	0.012	-0.001	15	0.012	-0.001	15
Pooled	0.020	-0.001	45	0.031	0.012	45	0.020	0.000	45

*Number of equating tables included in the pooled error estimates

It was seen earlier that linear methods were superior to equipercentile and STST methods for equating parallel composites directly. When nonparallel composites were equated, however, this was not true. That is, equipercentile equating outperformed linear equating (in terms of RMSE) for all instances of direct-composite equating; STST was only slightly worse than equipercentile equating for the direct power composites. For example, when power composites were equated using the smaller sample size, the pooled RMSE for the linear method was 0.031; the corresponding figures for the equipercentile and STST methods were 0.019 and 0.021, respectively. When AFQT composites were equated, these figures were 0.024 (linear) and 0.016 (equipercentile). Equating AFQT to power composites yielded linear and equipercentile RMSEs of 0.045 and 0.038, respectively. Bias was small throughout.

Using a selected examinee sample to directly equate composites resulted in large increases in both RMSE and bias for the conventional methods; STST was robust against this manipulation. Pooled RMSEs for the power and AFQT composites were 0.041 and 0.034, respectively; mean bias indices were 0.018 and 0.014. The pooled RMSE and bias for equating a power composite to an AFQT composite were 0.062 and 0.030, respectively. These pooled values reflect the effect of using a selected examinee sample when conventional equating methods were used; for STST, RMSE and bias were essentially unchanged.

When nonparallel composites were formed from equated subtests, no sample-size effect was evident for any of the equating methods or either of the two types of composites. When a selected sample was used for equating, however, bias increased from approximately zero to 0.022 for the conventional methods; STST was unbiased even when selected samples were used.

Discussion

Increasing the examinee sample size from 1,000 to 2,400 examinees had only a minor effect on the accuracy of composite equating. The use of a selected examinee sample caused an increase in equating error when nonparallel composites were directly equated, when unlike composites were directly equated, and when parallel AFQT composites were equated using equipercentile procedures. Strong true-score theory was unaffected by the use of selected examinee samples.

Test Lengths and Difficulties

Results

The error indices computed when parallel composites were equated using various levels of composite length and difficulty are presented in Table 36. When short composites were directly equated, it made little difference in equating accuracy (measured by RMSE) whether the items were easy or difficult. Biases were slightly positive for equating easy composites, slightly negative for difficult ones. The single exception to this bias pattern occurred when power composites were equated to AFQT composites. In this case, bias decreased from

0.012 to 0.007; there was little change in RMSE. For the long composites, the only nontrivial effect due to composite difficulty occurred when AFQT composites were directly equated. In that case, pooled RMSE decreased from 0.008 to 0.004. In general, the longer composites were equated with the same amount of error as were the shorter composites.

Table 36
True-Score Error Indices for Equating Parallel Composites Using Various Levels of Composite Length and Difficulty

	Composite difficulty					
	Easy			Difficult		
Composite length	RMSE	Bias	N*	RMSE	Bias	N*
Short composites						
Direct power	0.009	0.005	24	0.008	-0.002	24
Direct AFQT	0.008	0.004	12	0.006	-0.003	12
Direct power to AFQT	0.037	0.012	18	0.036	0.007	18
Equated power subtests	0.006	0.004	36	0.006	-0.003	36
Equated AFQT subtests	0.005	0.003	27	0.005	-0.003	27
Pooled	0.016	0.005	117	0.015	-0.001	117
Long composites						
Direct power	0.007	0.003	18	0.005	-0.001	18
Direct AFQT	0.008	0.002	12	0.004	-0.001	12
Direct power to AFQT	0.036	0.008	18	0.035	0.006	18
Equated power subtests	0.004	0.001	36	0.005	-0.001	36
Equated AFQT subtests	0.004	0.001	27	0.005	-0.001	27
Pooled	0.015	0.002	111	0.015	0.000	111

*Number of equating tables included in the pooled error estimates.

Table 37 presents the error indices computed when nonparallel composites were equated. As was observed when individual subtests were equated, the varying of item difficulty across composites being equated resulted in much larger equating errors than did the varying of composite length only; this was true for all types of composites investigated in this study. For example, when difficulty was varied across power composites that were to be directly equated, the resulting pooled RMSEs were equal to 0.036 and 0.037 for the short and long composites, respectively. Varying composite length across these same composites yielded pooled RMSEs of 0.013 and 0.011 for the easy and difficult composites, respectively. This same pattern of errors was evident for all types of composites.

Table 37

True-Score Error Indices for Equating Nonparallel Composites Using Various Levels of Composite Length and Difficulty

	Different difficulty				Different length				Different length and difficulty		N of equatings per cell
	Short		Long		Easy		Difficult				
Equating Method	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	
Direct power											
Linear	0.046	0.011	0.045	0.008	0.009	0.003	0.009	0.001	0.050	0.010	9
Equipercentile	0.033	0.014	0.026	0.008	0.016	0.005	0.012	0.001	0.035	0.012	9
STST	0.020	0.003	-	-	-	-	-	-	-	-	6
Pooled	0.036	0.010	0.037	0.008	0.013	0.004	0.011	0.001	0.043	0.011	18*
Direct AFQT											
Linear	0.036	0.009	0.035	0.006	0.006	0.000	0.007	-0.002	0.039	0.005	6
Equipercentile	0.023	0.008	0.024	0.006	0.017	0.004	0.009	-0.001	0.028	0.006	6
Pooled	0.031	0.008	0.030	0.006	0.013	0.002	0.008	-0.001	0.034	0.005	12
Direct power to AFQT											
Linear	0.061	0.017	0.059	0.012	0.035	0.008	0.037	0.007	0.062	0.014	9
Equipercentile	0.052	0.020	0.047	0.014	0.040	0.011	0.038	0.008	0.053	0.017	9
Pooled	0.056	0.019	0.053	0.013	0.038	0.009	0.038	0.007	0.058	0.015	18
Equated power subtests											
Linear	0.048	0.010	0.047	0.007	0.014	0.004	0.015	0.001	0.058	0.011	9
Equipercentile	0.037	0.016	0.028	0.010	0.016	0.003	0.016	-0.001	0.046	0.013	9
IRT	0.025	0.005	0.018	0.005	0.012	0.006	0.010	0.000	0.025	0.007	9
STST	0.019	-0.003	0.012	-0.001	0.008	0.003	0.010	0.000	0.019	-0.005	9
Pooled	0.034	0.007	0.029	0.005	0.013	0.004	0.013	0.000	0.040	0.007	36
Equated AFQT subtests											
Linear	0.038	0.007	0.037	0.005	0.012	0.003	0.012	0.001	0.046	0.008	9
Equipercentile	0.029	0.012	0.022	0.008	0.013	0.004	0.013	0.000	0.037	0.011	9
STST	0.016	-0.003	0.010	-0.001	0.007	0.002	0.008	0.000	0.015	-0.004	9
Pooled	0.029	0.005	0.026	0.004	0.011	0.003	0.011	0.000	0.035	0.005	27

*Except: N = 24 for first cell.

In general, equipercentile methods worked best when composite scores were directly vertically equated. For directly equating power composites, however, STST outperformed even the equipercentile methods. Equipercentile equating bias, however, was almost always greater than or equal to linear bias for the direct composites; STST was essentially unbiased. Conversely, linear methods performed best when composites of constant difficulty (but varying lengths) were directly equated.

When composites were formed from equated subtests, STST consistently outperformed all other equating methods; the conventional methods consistently yielded the largest errors.

Equating errors were largest when both length and difficulty were varied across the composites being equated. In this case, equipercentile methods worked best for the direct composites; STST worked best when composites were formed from equated subtests.

Discussion

Whenever parallel composites were equated, difficulty had only a minor effect on equating accuracy. In general, longer composites were equated with less error than were the shorter composites. This would be expected as longer tests are usually better estimates of ability than are shorter tests.

For nonparallel composites, varying difficulty across composites being equated resulted in much larger errors than did varying composite length. STST was shown to be best for the limited conditions under which it was applied. Equipercentile procedures were better than linear procedures for vertical equating.

Real-Data Application

Results

Table 38 presents observed-score error indices computed when the equating transformations were applied to the item response data from an independent sample of 1,000 examinees. Table 38 indicates that nearly all the combinations of equating methods and data collection designs yielded equating transformations that contained identical amounts of error; there were only a few exceptions.

The data set contained responses to two randomly parallel Word Knowledge subtests. Given the results from the simulated parallel subtests that were reported earlier, one would expect the linear equating method to perform at least as well as any of the more complex

Table 38
Observed-Score Error Indices for Equating Methods and Data Collection Designs: Real-Data Verification

	Single group		Equivalent groups		Anchor test	
Equating method	RMSE	Bias	RMSE	Bias	RMSE	Bias
Linear	0.067	0.001	0.066	-0.002	0.067	0.002
Equipercentile	0.068	-0.002	0.065	-0.005	0.069	-0.002
IRT	0.065	-0.003	0.067	-0.011	0.065	0.001
STST	0.068	0.004	0.069	-0.005	0.080	-0.003

equating methods. This was, in fact, what was observed. In general, the conventional equating methods performed about as well as did any of the other, more complex methods for this case of parallel-test equating. The linear equating method typically resulted in less bias than did any of the other methods; linear bias never exceeded 0.002 in absolute value, whereas the other methods resulted in bias that ranged from 0.001 to 0.011 in absolute value. Linear RMSEs (0.066-0.067) were about the same level as those from the equipercentile (0.065-0.069) and IRT (0.065-0.067) methods, and were slightly smaller than those from STST (0.068-0.080).

IRT equating using the equivalent-groups design yielded an unexpectedly large value of bias (when compared to the other methods and designs); the bias index for this design was -0.011. The IRT RMSE was higher for the equivalent-groups design than for the other two designs.

Strong true-score theory yielded moderate (comparatively speaking) levels of bias for all three designs (0.003 to 0.005 in absolute value). The RMSE for the anchor-test design (0.080) was larger for STST than it was for any other value in the table.

The standard error of the difference between the equated scores and the observed scores was computed for this data set. This standard error is an estimate of the measurement error in these difference measures that would be expected in the absence of any explicit equating error. It was computed using an estimate of the test's reliability (coefficient alpha) that is a lower-bound estimate of the actual reliability. As such, the estimate of this standard error is an upper-bound estimate of the actual standard error. The standard error for this data set was equal to 0.068.

The observed-score RMSEs reported in Table 38 were typically no larger than the estimated standard error of the difference between the equated and observed scores. This suggests that the error involved in estimating the standard error was at least as large as the equating error itself. Partialing measurement error out of the observed-score RMSEs is, thus, not a feasible means of estimating equating error involved in real data.

Discussion

It is evident from these analyses that equating error and corresponding differences among the equating methods are obscured by the comparatively large standard errors present in real data. Equating error, especially for parallel tests, would be expected to be fairly small in magnitude: A difference as large as half a score point would translate to a value less than 0.02 on the proportion-correct metric for a 30-item test; most of the RMSEs computed when parallel subtests were simulated and equated were less than half that size.

On the other hand, the standard error of the difference between equated and observed scores was computed to be 0.068. Even when one considers that this is an upper-bound estimate of the actual standard error, it is obvious that equating error is easily overwhelmed by the amount of measurement error in the data. The net effect of this phenomenon was to make the criteria of equating accuracy (functions of the difference between equated and observed scores) insensitive to all but very large amounts of equating error. This, in turn, suggests that the criteria are insensitive to relatively small differences across equating methods. This problem cannot be readily solved by partialing measurement error out of the observed-score RMSEs, given that the error involved in estimating the standard error is probably as large as the equating error itself.

Only when a test is equated to itself (directly or through a chain of other tests) can meaningful interpretations concerning equating accuracy be made from real data. In this case, there is a criterion for evaluating equating accuracy that involves only the equating transformation and is thus independent of examinee responses: Each test score should be equated to itself, and any deviation of the observed transformation from this "identity" transformation constitutes an equating error. In all other instances, researchers should exercise caution when interpreting results from equating studies that rely on observed-score criteria.

CONCLUSIONS AND RECOMMENDATIONS

Individual Subtests

Smoothing Methods

For the situations simulated here, none of the smoothing methods yielded an equating transformation that was more accurate than that yielded by "no smoothing." It is not clear whether other testing situations or other implementations of smoothing procedures would have yielded different results.

Equating Methods

Theoretically, the conventional (linear and equipercentile) and STST equating methods are appropriate whenever parallel subtests are to be equated; IRT methods are also appropriate if the subtests to be equated are unidimensional and not speeded. Previous research has indicated that the conventional and IRT procedures yield essentially the same results in these conditions. No studies have investigated the utility of STST equating procedures.

Only STST procedures are theoretically appropriate for equating nonparallel subtests in every situation; IRT procedures are appropriate for unidimensional power subtests. Studies comparing conventional and IRT methods for equating subtests of equal difficulty have yielded equivocal results. Conventional equating methods have not been found adequate for vertical equating situations (i.e., for equating subtests of different difficulties). Studies indicate that a pseudo-guessing parameter needs to be incorporated in any IRT model that is to be used for vertical equating.

The present study found that complex equating methods (such as IRT or STST) need not be used when parallel subtests are equated. The simpler conventional equating methods performed just as well as, and usually better than, the more complex methods for equating parallel tests. Either linear or equipercentile methods are recommended for parallel power tests. However, when nonparallel power subtests are equated, the conventional methods fail to perform adequately; IRT and STST methods are clearly better for this case. Linear equating performed best for parallel and nonparallel speeded tests. In general, nonparallel subtests were equated with greater error than were parallel subtests.

Data Collection Designs

All data collection designs performed adequately for equating parallel power tests. For equating nonparallel tests by IRT methods or for equating speeded tests, however, the single-group design is clearly preferable and should be used where practically feasible. The equivalent-groups design should be used to equate tests only when the two examinee samples are, in fact, equivalent in ability. Whenever power tests are equated using samples that differ in ability level, the anchor-test design is essential; combining it with the equipercentile equating method is advisable.

Sample Sizes

In most of the published studies of test equating procedures, data were obtained from national testing programs with very large numbers of examinees (typically several thousand); hence, sample size was not an issue in these studies. Yen's (1982) sample-size manipulation of 1,000 vs. 2,000 had no effect on the accuracy of equipercentile equating. Douglass (1980, 1981) varied sample size from 200 to 800 examinees and found that this manipulation influenced the consistency of two-parameter-IRT equatings but was not a salient factor for the one-parameter model. Similarly, Kolen and Whitney (1982) suggested that a sample of 200 examinees was not large enough for adequate equating using the three-parameter IRT model.

The present study confirmed Yen's finding; that is, there was little advantage to be gained by increasing the sample size from 1,000 to 2,400 examinees; how small the sample size can get before equating accuracy is markedly affected cannot be determined from these data. Parallel subtests can be adequately equated with a selected examinee sample. When nonparallel subtests are equated, a selected examinee sample should not be used in conjunction with the conventional equating methods. IRT equating was only slightly affected by using a selected examinee sample, and strong true-score theory appears robust against this manipulation.

Test Lengths and Difficulties

No studies to date have provided information concerning the minimum number of items a test must contain before equating procedures can be appropriately implemented. The only information concerning item difficulties arises from the literature on vertical equating, where tests of unequal difficulty are equated for later administration to examinees of unequal ability.

Theoretically, only the STST method is appropriate for equating all nonparallel tests (i.e., tests of unequal difficulty and/or reliability); IRT methods are also appropriate for unidimensional power subtests. Previous studies suggest that vertical equating can be successful if IRT methods (where the IRT model incorporates some provision for guessing) are used and the groups do not vary widely in ability.

In this study, equating accuracy (defined by RMSE) was not markedly affected when subtest length was doubled (at least for the test lengths of 15 and 25 investigated here), nor did it matter whether easy or difficult subtests were equated. However, scores on easy subtests may be overestimated while scores on difficult subtests may be underestimated, at least if short parallel tests are equated. Accuracy was not affected when subtests of different lengths were equated.

When the difficulty level varies across the subtests being equated, conventional equating methods should not be used. IRT or STST methods should be used for these vertical equating situations.

Composites

Equated composite scores can be defined and constructed in any of several different ways. For example, composite scores can first be computed as the weighted sum of individual (unequated) subtest scores. These composite scores can then be directly equated using conventional or STST methods; because item response theory assumes that each test score is unidimensional, IRT equating methods are not applicable in this case. This direct-equating method is usually considered to be the preferred method of equating composite scores because it arises so naturally from the goal of composite-score equating: to define equivalent scores on two composites of subtests.

Alternatively, the individual subtests can first be separately equated by the conventional, STST, or (where appropriate) IRT methods. Equated composite scores can then be formed for future examinees by applying the composite weights to their equated subtest scores. It is necessary to use this procedure when each group of examinees is administered only a single subtest and composite scores cannot be directly equated. In this case, a separate transformation table must be constructed and applied for each subtest in the composite.

Composite scores can also be equated indirectly using conventional linear procedures that take into account the original composite weights, subtest means and standard deviations, and the intercorrelations among the subtest scores. This procedure is actually a reformulation of the (direct) linear equating model in which two composite scores are considered to be equated if their corresponding standard scores are equal. The advantage of using this indirect procedure is that composite scores can be equated even if all examinees do not take all the subtests in a battery. With partial data, then, this procedure becomes an approximation to the procedure described above for equating composite scores directly.

No previous study systematically investigated the merits of the various procedures for equating composites of test scores. This study found few practical differences among the procedures investigated for equating test composites.

In general, the equating of parallel composites is most successful when individual subtests are first equated and composites are formed from the equated subtests; there are essentially no differences among the equating methods for this case. If parallel subtests are to be directly equated, then linear methods should be used because they yield smaller errors.

Nonparallel composites are best equated by forming composites from subtests that have been previously equated using IRT or STST methods. In general, STST is the preferred method for equating nonparallel composites in all those conditions where it is practicable (e.g., shorter test lengths, large sample sizes, etc.). If conventional equating methods need to be used, the nonparallel composites should be directly equated and equipercentile procedures should be employed. If the goal of the equating procedure is to yield unbiased equated scores at the expense of all other types of error, then the STST method should be used in all cases.

Parallel composites were equated with less error than were nonparallel composites (as was the case for individual subtests). Composites composed of different subtests (e.g., power and AFQT composites or, worse, composites with no subtests in common) should not be equated; this type of composite equating is inappropriate both theoretically and practically.

"Indirect" composite-equating procedures, where composite scores are (linearly) equated by using subtest statistics and intercorrelations, can be a good substitute for the direct linear equating of composite scores when examinee response data are not available for all subtests in a battery.

No clear recommendations can be made regarding the choice of a data collection design or sample size for equating test composites, since no consistent differences among designs and sizes were noted. Selected examinee samples should not be used to equate anything other than direct parallel power and AFQT composites. Vertical equating of composites is not recommended.

REFERENCES

Angoff, W. H. (1962). Scales with nonmeaningful origins and units of measurement. *Educational and Psychological Measurement*, *22*, 27-34.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington: American Council on Education.

Assessment Systems Corporation. (1982). *The Minnesota computerized adaptive testing system* [Computer program manual]. St. Paul, MN: Author.

Beard, J. G., & Pettie, A. L. (1979, April). *A comparison of linear and Rasch equating results for basic skills assessment tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bejar, I. I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. 81-3, Educational Testing Service Report No. 81-35). New York: College Entrance Examination Board, 1981.

Bejar, I. I., & Wingersky, M. S. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement*, *6*, 309-325.

Bianchini, J. C., & Loret, P. G. (1974). *Anchor Test Study: Final Report*. Washington: U.S. Department of Health, Education, and Welfare, U.S. Office of Education.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical Theories of Mental Test Scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, *29*, 610-611.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Ed.), *Test equating* (pp. 9-49). New York: Academic Press.

Brokaw, L. D., & Burgess, G. G. (1957). Development of Airman Classification Battery AC-2A (AFPTRC-TR-57-1, AD-131 422). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 175-195). Vancouver: Educational Research Institute of British Columbia.

Cook, L. L., Eignor, D. R., & Petersen, N. S. (1982, April). A study of the temporal stability of IRT item parameter estimates. Paper presented at the annual meeting of the American Educational Association, New York.

Dailey, J. T., Shaycoft, M. F., & Orr, D. B. (1962). Calibration of Air Force selection tests to Project TALENT norms (PRL-TDR-62-6, AD-285 185). Lackland Air Force Base, TX: 6570th Personnel Research Laboratory.

Divgi, D. R. (1980, April). A nonparametric test for comparing goodness of fit in latent trait theory. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Divgi, D. R. (1981a, April). Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Divgi, D. R. (1981b). Model-free evaluation of equating and scaling. Applied Psychological Measurement, 5, 203-208.

Douglass, J. B. (1980, April). Applying latent trait theory to a classroom examination system: Model comparison and selection. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Douglass, J. B. (1981, April). A comparison of item response theory models for use in a classroom examination system. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Edwards, D. S., & Hahn, C. P. (1962). Development of Airman Qualifying Examination-62 (PRL-TDR-62-7, AD-284 775). Lackland Air Force Base, TX: 6570th Personnel Research Laboratory.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), Educational Measurement. Washington: American Council on Education.

Fleishman, A. I. (1978). A method for simulating nonnormal distributions. Psychometrika, 43, 521-532.

Garcia-Quintana, R. A., & Johnson, L. M. (1979, April). *Equating two forms of a criterion-referenced test by using norm-referenced data: An illustration of two methods*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Golub-Smith, M. (1980, April). *The application of Rasch model equating techniques to the problem of interpreting longitudinal performance on minimum competency tests*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw-Hill.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating. *Applied Psychological Measurement*, *5*, 187-201.

Gustafsson, J-E. (1979). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement*, *16*, 153-158.

Hicks, M. M. (1983). True score equating by fixed *b*'s scaling: A flexible and stable equating alternative. *Applied Psychological Measurement*, *7*, 255-266.

Holmes, S. E. (1981, June). *Vertical equating with the Rasch model*. Paper presented at the spring conference of the Washington Educational Research Association.

Holmes, S. E. (1982a, March). *The effects of test content match and number of items on the accuracy of trait estimates from tests equated with the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Holmes, S. E. (1982b). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement*, *19*, 139-147.

Jaeger, R. M. (1981). Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement*, *18*, 23-28.

Jensema, C. J. (1976). A simple technique for estimating latent trait mental parameters. *Educational and Psychological Measurement*, *36*, 705-715.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18*, 1-11.

Kolen, M. J. (1983). Effectiveness of analytic smoothing in equipercentile equating (ACT Technical Bulletin No. 41). Iowa City: American College Testing Program.

Kolen, M. J. & Whitney, D. R. (1982). Comparison of four procedures for equating the Tests of General Educational Development. Journal of Educational Measurement, 19, 279-293.

Linn, R. L. (1975). Anchor Test Study: The long and the short of it. Journal of Educational Measurement, 12, 201-214.

Lindgren, B. W. (1976). Statistical theory (3rd ed.). New York: Macmillan.

Lord, F. M. (1950). Notes on comparable scales for test scores (Research Bulletin 50-48). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1965). A strong true-score theory, with applications. Psychometrika, 30, 239-270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 34, 259-299.

Lord, F. M. (1975). Automated hypothesis tests and standard errors for nonstandard problems. The American Statistician, 29, 56-59.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1981a). The standard error of equipercentile equating (Research Report 81-48). Princeton: Educational Testing Service.

Lord, F. M. (1981b). Standard error of an equating by Item Response Theory (Research Report 81-49). Princeton: Educational Testing Service.

Lord, F. M. (1982a). Item response theory and equating - a technical summary. In P. W. Holland & D. B. Rubin (Ed.), Test Equating (pp. 141-148). New York: Academic Press.

Lord, F. M. (1982b). Standard error of an equating by item response theory. Applied Psychological Measurement, 6, 463-472.

Lord, F. M. (1982c). The standard error of equipercentile equating. Journal of Educational Statistics, 7, 165-174.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1983). Comparison of IRT observed-score and true-score 'equatings' (Research Report RR-83-26-ONR). Princeton, NJ: Educational Testing Service.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Madden, H. L., & Lecznar, W. B. (1965). Development and standardization of Airman Qualifying Examination-64 (PRL-TR-65-14, AD-622 807). Lackland Air Force Base, TX: Personnel Research Laboratory.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 14, 139-160.

Marco, G. L., Petersen, N., & Stewart, E. (1980). A test of the adequacy of curvilinear score equating models. Proceedings of the 1979 Computerized Adaptive Testing Conference (pp. 167-196). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Marks, E., & Lindsay, C. A. (1972). Some results relating to test equating under relaxed test form equivalence. Journal of Educational Measurement, 9, 45-56.

Modu, C. C. (1982, April). The robustness of latent trait models for achievement test score equating. Paper presented at the annual meeting of the American Educational Research Association, New York.

Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Ed.), Test equating (pp. 169-191). New York: Academic Press.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Phillips, S. E. (1983). Comparison of equipercentile and item response theory equating when the scaling test method is applied to a multilevel achievement battery. Applied Psychological Measurement, 7, 267-281.

Pieters, J. P. M., & van der Ven, A. H. G. S. (1982). Precision, speed, and distraction in time-limit tests. Applied Psychological Measurement, 6, 93-109.

Pike, M. C., & Hill, I. D. (1966). Algorithm 291: Logarithm of gamma function. *Communications of the ACM*, *9*, 684.

Prestwood, J. S., Vale, C. D., Massey, R. H., and Welsh, J. R. (in press). The development of an adaptive item pool for the ASVAB. Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Ree, M. J. (1981). The effects of item calibration sample size and item pool size on adaptive testing. *Applied Psychological Measurement*, *5*, 11-19.

Ree, M. J., Mathews, J. J., Mullins, C. J., & Massey, R. H. (1982). *Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10* (AFHRL-TR-81-49, AD-A114 714). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik*, *10*, 177-183.

Rentz, R. R., & Bashaw, W. L. (1975). *Equating reading tests with the Rasch model: Final Report*. Athens, GA: University of Georgia, Educational Research Laboratory.

Rentz, R. R., & Bashaw, W. L. (1977). The National Reference Scale for reading: An application of the Rasch model. *Journal of Educational Measurement*, *14*, 161-179.

Rubin, D. B. (1982). Discussion of "Observed-score test equating: A mathematical analysis of some ETS equating procedures." In P. W. Holland & D. B. Rubin (Ed.), *Test equating* (pp. 51-54). New York: Academic Press.

Slinde, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? *Journal of Educational Measurement*, *14*, 23-32.

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, *15*, 23-35.

Slinde, J. A., & Linn, R. L. (1979a). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement*, *16*, 159-165.

Slinde, J. A., & Linn, R. L. (1979b). The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement*, *3*, 437-452.

Stock, W. A., Kagan, D. M., & Van Wagenen, R. K. (1980). Graduate Record Examination and Miller Analogies Test scores: Examining four methods of equivalencing. Educational and Psychological Measurement. 4, 829-834.

Stocking, M., Wingersky, M. S., Lees, D. M., Lennon, V., & Lord, F. M. (1973). A program for estimating the relative efficiency of tests at various ability levels, for equating true scores, and for predicting bivariate distributions of observed scores (Research Memorandum 73-24). Princeton, NJ: Educational Testing Service.

Sympson, J. B. (1982, June). Item calibrations for computerized adaptive testing (CAT) prototype item pools. Paper presented at the 1982 Computerized Adaptive Testing Conference, Minneapolis.

Thompson, C. A. (1958). Development of the Airman Qualifying Examination Forms D and E, Part I (WADC-TR-58-94(I), AD-151 045). Lackland Air Force Base, TX: Wright Air Development Center.

Vale, C. D, & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. Psychometrika, 48, 465-471.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). Methods for linking item parameters (AFHRL-TR-81-10, AD-A105 508). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Vitola, B. M., & Alley, W. E. (1968). Development and standardization of Air Force composites for the Armed Services Vocational Aptitude Battery (AFHRL-TR-68-110, AD-688 470). Lackland Air Force Base, TX: Air Force Human Resources Laboratory.

Weeks, J. L., Mullins, C. J., & Vitola, B. M. (1975). Airman classification batteries from 1948 to 1975: A review and evaluation (AFHRL-TR-75-78, AD-A026 470). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Wichmann, B. A., & Hill, I. D. (1982). An efficient and portable pseudo-random number generator. Applied Statistics, 31, 188-190.

Wingersky, M. S., Lees, D. M., Lennon, V., & Lord, F. M. (1969). A computer program for estimating true-score distributions and graduating observed-score distributions (Research Memorandum 69-4). Princeton, NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Journal of Educational Measurement*, *29*, 23-48.

Yen, W. M. (1982, April). *Obtaining some degree of correspondence between unequatable scores*. Paper presented at the annual meeting of the American Educational Research Association, New York.

APPENDIX A. STRONG TRUE-SCORE THEORY

Grouping Scores into Intervals

The purpose of grouping is to make the sample frequency distribution smoother. This procedure attempted to group scores such that no cell frequency was small and each had about 1/(number of groups) of the cases.

The following loop was repeated until one of the following reasonable arbitrary limits was reached: (a) the number of groups reached the maximum of 25; (b) the size of the smallest group was less than 25; or (c) the variable "portion" (originally set to 1.0 and reduced by 0.1 on each loop) was less than .199.

(1) Alternate
  (a) forming the next lowest group (group 1 being the first) by combining the next available low scores (having started with 0, 1, 2,...) until their combined sample size exceeds (portion *N/25), and
  (b) forming the next highest group (group 25 being the first) by combining the next available high scores (having started with n, (n-1), (n-2), ...) until their combined sample size exceeds (portion *N/25)
  until there are no ungrouped scores (in which case any blank groups in the middle are eliminated, decreasing the total number of groups) or there are no groups left to put the remaining scores into (in which case the remaining scores are divided between the two middle groups).
(2) Compute group frequencies.
(3) Find the size of the smallest group.
(4) Decrease portion by 0.1.

Computational Formulas for the Constants a_{xu}

The constants, a_{xu}, in the strong true-score theory general model are by definition

$$a_{xu} = \sum_{y:u} a_{xy}$$

where $a_{xy} = \int_0^1 \gamma(\zeta)\ h(x|\zeta)\ h(y|\zeta)\ \partial\zeta$. [20]

Computational formulas for each element of the integral must be found, and then the integral must be evaluated.

First, let $\gamma(\zeta) = 1$, a smooth density function.

Then, from Lord (1965, Equations 5-7 with $r = 2$),

$$h(x|\zeta) = \binom{n}{x} \zeta^{x} (1-\zeta)^{n-x} + k\ \zeta(1-\zeta) \cdot \left\{ -\binom{n-2}{x} \zeta^{x} (1-\zeta)^{n-x-2} + 2\binom{n-2}{x-1} \cdot \zeta^{x-1} (1-\zeta)^{n-x-1} - \binom{n-2}{x-2} \zeta^{x-2} (1-\zeta)^{n-x} \right\}$$

where $k \doteq \dfrac{n^2 (n-1)\ s_p}{2\{n^2 \overline{pq} - s_x^2 - n s_p^2\}}$ [Lord, 1965, Equation 46]

p = conventional item difficulty, and
$q = 1 - p$.

This can be simplified to

$$\begin{aligned} h(x|\zeta) &= \binom{n}{x} \zeta^{x} (1-\zeta)^{n-x} - k \binom{n-2}{x} \zeta^{x+1} (1-\zeta)^{n-x-1} + 2k \binom{n-2}{x-1} \zeta^{x} \cdot \\ &\quad (1-\zeta)^{n-x} - k \binom{n-2}{x-2} \zeta^{x-1} (1-\zeta)^{n-x+1} \\ &= \left\{ \binom{n}{x} + 2k \binom{n-2}{x-1} \right\} \zeta^{x} (1-\zeta)^{n-x} \\ &\quad - k \binom{n-2}{x} \zeta^{x+1} (1-\zeta)^{n-x-1} \\ &\quad - k \binom{n-2}{x-2} \zeta^{x-1} (1-\zeta)^{n-x+1} . \end{aligned}$$

To simplify the notation,

let $w_{x1} = \{\binom{n}{x} + 2k\binom{n-2}{x-1}\}$

$w_{x2} = -k\binom{n-2}{x}$

$w_{x3} = -k\binom{n-2}{x-2}$.

Then, the product of the two h functions becomes

$$
\begin{aligned}
h(x|\zeta)\, h(y|\zeta) &= w_{x1}\, w_{y1}\, \zeta^{x+y}\, (1-\zeta)^{2n-x-y} \\
&+ w_{x2}\, w_{y1}\, \zeta^{x+y+1}\, (1-\zeta)^{2n-x-y-1} \\
&+ w_{x3}\, w_{y1}\, \zeta^{x+y-1}\, (1-\zeta)^{2n-x-y+1} \\
&+ w_{x1}\, w_{y2}\, \zeta^{x+y+1}\, (1-\zeta)^{2n-x-y-1} \\
&+ w_{x2}\, w_{y2}\, \zeta^{x+y+2}\, (1-\zeta)^{2n-x-y-2} \\
&+ w_{x3}\, w_{y2}\, \zeta^{x+y}\, (1-\zeta)^{2n-x-y} \\
&+ w_{x1}\, w_{y3}\, \zeta^{x+y-1}\, (1-\zeta)^{2n-x-y+1} \\
&+ w_{x2}\, w_{y3}\, \zeta^{x+y}\, (1-\zeta)^{2n-x-y} \\
&+ w_{x3}\, w_{y3}\, \zeta^{x+y-2}\, (1-\zeta)^{2n-x-y+2} \\
&= (w_{x1}\, w_{y1} + w_{x2}\, w_{y3} + w_{x3}\, w_{y2})\, \zeta^{x+y}\, (1-\zeta)^{2n-x-y} \\
&+ (w_{x1}\, w_{y2} + w_{x2}\, w_{y1})\, \zeta^{x+y+1}\, (1-\zeta)^{2n-x-y-1} \\
&+ (w_{x1}\, w_{y3} + w_{x3}\, w_{y1})\, \zeta^{x+y-1}\, (1-\zeta)^{2n-x-y+1} \\
&+ w_{x2}\, w_{y2}\, \zeta^{x+y+2}\, (1-\zeta)^{2n-x-y-2} \\
&+ w_{x3}\, w_{y3}\, \zeta^{x+y-2}\, (1-\zeta)^{2n-x-y+2} .
\end{aligned}
$$

A Beta function (Lindgren, 1976, pp. 328-9) is defined as

$$\beta\ (x+1,\ y+1) = \int_0^1 \zeta^x\ (1-\zeta)^y\ \partial\zeta$$

The Beta function is also equal to a product of Gamma functions

$$\beta\ (s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+k)}$$

and there is an algorithm available (Pike & Hill, 1966, algorithm 291) to solve for Γ. Thus, the Beta function can be used to evaluate the integral.

Substituting quantities derived above into Equation 20

$$a_{xy} = \int_0^1 \gamma(\zeta)\ h(x|\zeta)\ h(y|\zeta)\ \partial\zeta$$

$$= \int_0^1 h(x|\zeta)\ h(y|\zeta)\ \partial\zeta$$

$$= (w_{x1}\ w_{y1} + w_{x2}\ w_{y3} + w_{x3}\ w_{y2})\ \beta\ (x+y+1, 2n-x-y+1)$$

$$+ (w_{x1}\ w_{y2} + w_{x2}\ w_{y1})\ \beta\ (x+y+2, 2n-x-y)$$

$$+ (w_{x1}\ w_{y3} + w_{x3}\ w_{y1})\ \beta\ (x+y,\ 2n-x-y+2)$$

$$+ (w_{x2}\ w_{y2})\ \beta\ (x+y+3, 2n-x-y-1)$$

$$+ (w_{x3}\ w_{y3})\ \beta\ (x+y-1, 2n-x-y+3)$$

for $x = 1, 2, \ldots, n$, and

$y = 1, 2, \ldots, n$.

Rescaling $\hat{\lambda}_u$'s

$\hat{\lambda}_u$'s must be rescaled so that they are all nonnegative to guarantee $\hat{g}(\zeta)$ nonnegative for $0 \leq \zeta \leq 1$. This was accomplished by setting all negative $\hat{\lambda}_u$'s to a tenth the size of the smallest positive $\hat{\lambda}_u$ and then dividing all $\hat{\lambda}$'s by their sum because (Lord, 1969, Equation 31)

$$\sum_{u=1}^{U} A_u \lambda_u = \sum_{x=0}^{n} \phi(x) = 1$$

where $A_u = \sum_{x=0}^{n} a_{xu}$.

Maximum Likelihood Procedures to Refine the $\hat{\lambda}_u$'s

(1) Compute $\hat{\phi}(x)$ (from Equation 20 with $\hat{\lambda}$'s inserted) and $\hat{\phi}_u'(x)$ for $\underline{x} = 0, 1, \ldots, n$ and $\underline{u} = 1, 2, \ldots \underline{U}$. The $\hat{\phi}_u'(x)$ are functions of the a_{xu}'s.

$$\phi(x) = \sum_{u=1}^{U} a_{xu} \lambda_u$$

$$= \sum_{u=1}^{U-1} a_{xu} \lambda_u + a_{xU} \lambda_U \,. \qquad [20]$$

$$\sum_{u=1}^{U} A_u \lambda_u = \sum_{u=1}^{U-1} A_u \lambda_u + A_U \lambda_U = 1 \,. \qquad \text{[Lord, 1969, Equation 31]}$$

Thus $\lambda_u = (1 - \sum_{u=1}^{U-1} A_u \lambda_u)/A_U$, and

$$\phi(x) = \sum_{u=1}^{U-1} a_{xu} \lambda_u + a_{xu} \, (1 - \sum_{u=1}^{U-1} A_u \lambda_u)/A_U$$

$$= a_{xu}/A_U + \sum_{u=1}^{U-1} \lambda_u \, (a_{xu} - a_{xU} A_u/A_U) \,.$$

$$\phi'(x) = \frac{\partial \phi(x)}{\partial \lambda u} = a_{xu} - a_{xU} A_u/A_U$$

(2) Find $\frac{\partial \ln L}{\partial \lambda u}$, the first derivative of the log of the likelihood with respect to λ for $\underline{u} = 1, 2, \ldots, U$, derived as follows:

$$\ln L = \sum_{x=0}^{n} f_x \ln (\phi(x)) \,, \text{ and}$$

$$\frac{\partial \ln L}{\partial \lambda u} = \sum_{x=0}^{n} f_x \frac{1}{\phi(x)} \frac{\partial \phi(x)}{\partial \lambda u} = \sum_{x=0}^{n} f_x \frac{\phi'(x)}{\phi(x)} \,.$$

(3) Find $-\dfrac{\partial^2 \ln L}{\partial\lambda_u \partial\lambda_z}$, the negative of the second derivative of the log likelihood for pairs of nonzero $\hat{\lambda}$'s only for u, z = 1, 2, ..., (U-1). (There are only (U-1) independent $\hat{\lambda}$'s due to Lord (1969, Equation 31). The formula is derived as follows:

$$\frac{\partial^2 \ln L}{\partial\lambda_u^2} = \sum_{x=0}^{n} f_x \frac{\phi(x)\, \phi''(x) - \phi'(x)\, \phi'(x)}{\phi^2(x)} .$$

$$\phi''(x) = \frac{\partial\phi'(x)}{\partial\lambda_u} = 0 .$$

Thus, $$\frac{\partial^2 \ln L}{\partial\lambda_u^2} = \sum_{x=0}^{n} f_x \frac{-(\phi'(x))^2}{\phi^2(x)} = -\sum_{x=0}^{n} f_x \left[\frac{\phi'(x)}{\phi(x)}\right]^2 .$$

$$\frac{\partial^2 \ln L}{\partial\lambda_u \partial\lambda_z} = \sum_{x=0}^{n} f_x \frac{\phi(x)\dfrac{\partial^2\phi(x)}{\partial\lambda_u\partial\lambda_z} - \dfrac{\partial\phi(x)}{\partial\lambda_u}\cdot\dfrac{\partial\phi(x)}{\partial\lambda_z}}{\phi^2(x)}$$

$$= \sum_{x=0}^{n} f_x \left(\frac{-\dfrac{\partial\phi(x)}{\partial\lambda_u}\dfrac{\partial\phi(x)}{\partial\lambda_z}}{\phi^2(x)} \right)$$

$$= -\sum_{x=0}^{n} f_x \frac{\phi'_u(x)\, \phi'_z(x)}{\phi^2(x)} .$$

(4) Find the change values to be added to the old $\hat{\lambda}_u$'s by the Newton-Raphson iterative procedure, namely the

$$\delta_u = (\ln L)'/(-\ln L)'' \text{ for } \underline{u} = 1, 2, \ldots, (U-1).$$

(5) Compute the new $\hat{\lambda}_u$'s.

$$\text{new } \hat{\lambda}_u = \text{old } \hat{\lambda}_u + \delta_u \text{ for } \underline{u} = 1, 2, \ldots, (U-1).$$

(6) Find δ_u and new $\hat{\lambda}_u$ from Lord (1969, Equation 31) above.

(7) Set negative $\hat{\lambda}_u$'s to zero and rescale the remaining $\hat{\lambda}_u$'s so that the above equation obtains; recompute the δ_u's.

(8) Check whether convergence has occurred (largest δ_u < criterion value, 0.1 here) or the maximum number of loop has been reached (200, here); if neither condition is true, go back to step (1).

(9) Lambdas which had been set to zero were reinserted one at a time and the above refinement looping (steps 1-8) was repeated until none of the lambdas in the set were changed from zero during a cycle from $\hat{\lambda}_o$ through $\hat{\lambda}_u$.

Obtaining Estimated True Percentile for a Given True Score

The estimated true percentile for true score t is

$$\int_0^t g(\zeta)\, d\zeta \qquad [25]$$

A computational formula must be substituted for the $g(\zeta)$ and then the integral must be evaluated. Lord has shown that

$$g(\zeta) = \sum_{u=1}^{U} \lambda_u H_u(\zeta) \qquad \text{[Lord, 1969, Equation 40]}$$

and

$$H_u(\zeta) = \gamma(\zeta) \sum_{x:u} h(x|\zeta) = \sum_{x:u} \gamma(\zeta)\, h(x|\zeta). \qquad \text{[Lord, 1969, Equation 22]}$$

A Beta distribution was used for the frequency distribution so that combinations and factorials would be defined:

Let $\gamma(\zeta) = \zeta^{\alpha}(1-\zeta)^{\delta}$, with $\alpha = \delta = 2$.

From the derivation for a_{xu} above, we know that

$$h(x|\zeta) = w_1\, \zeta^{x}(1-\zeta)^{n-x} + w_2\, \zeta^{x+1}(1-\zeta)^{n-x-1} + w_3\, \zeta^{x-1}(1-\zeta)^{n-x+1}$$

where

$$w_1 = \binom{n}{x} + 2k\binom{n-2}{x-1},$$

$$w_2 = -k\binom{n-2}{x}, \text{ and}$$

$$w_3 = -k\binom{n-2}{x-2}.$$

Thus,

$$H_u(\zeta) = \sum_{x:u} \{w_1 \zeta^{x+\alpha}(1-\zeta)^{n-x+\delta}$$

$$+ w_2\, \zeta^{x+1+\alpha}(1-\zeta)^{n-x-1+\delta}$$

$$+ w_3\, \zeta^{x-1+\alpha}(1-\zeta)^{n-x+1+\delta}\}.$$

Substituting back into the integral and using the Beta function to evaluate the integral, as explained above in deriving computational formulas for a_{xu},

$$\int_0^1 g(\zeta)\partial\zeta = \sum_{u=1}^{U} \lambda_u \int_0^1 H_u(\zeta)\partial\zeta$$

$$= \sum_{u=1}^{U} \lambda_u \sum_{x:u} \{w_1 \beta(x+\alpha+1, n-x+\delta+1)$$

$$+ w_2 \beta(x+\alpha+2, n-x+\delta)$$

$$+ w_3 \beta(x+\alpha, n-x+\delta+2)\} = 1. \qquad \text{(because true scores are bounded by 0 and 1)}$$

To change the upper bound on the interval from 1 to t, the incomplete Beta function was substituted for the Beta function, yielding the

$$\text{proportion below true score } t = \int_0^t g(\zeta)\,\partial\zeta\,.$$

Initial Estimate of an Equated Score

A table of estimated true percentiles for scores on the old test was generated. If the estimated true percentile for the new-test score to be equated fell in the part of the table for which STST equating was possible, the initial value for the equated old-test score was the old-test score which defined the lower bound of the score-interval containing that true percentile.

Newton-Raphson Refinement of Equated Scores

Repeat the following procedure until $|\Delta| < 0.0001$ or ten iterations have been completed:

$$f = \int_0^{t_{old}} g(\zeta)\partial\zeta - p_{new}$$

where p_{new} = estimated true percentile on new test to be matched as closely as possible by the estimated true percentile on the equated score, and

t_{old} = the current value for the equated true score.

$$f' = g(t_{old}) - g(0) = g\ (t_{old}).$$

$$\Delta = f/f',$$

with Δ limited by $[-0.04, +0.05]$.

Updated $t_{old} = t_{old} - \Delta$.

☆U.S. GOVERNMENT PRINTING OFFICE 1984-569-053/453